THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Advancing Housing Image Classification Via Generated Images

American Family Insurance (AmFam), a Fortune 500 private insurance company based in the Midwest, requested the University of Chicago's Data Science Clinic to devise a model capable of classifying a house's architectural style from its image. This initiative was important for AmFam because architectural styles may influence the cost of reconstruction in the case of an insurance claim.

One major issue faced by the team and AmFam when developing the classification model is the problem with unbalanced data. Specifically, there were architectural styles that didn't have enough data to build an accurate classifier. The team decided to leverage state-of-the-art generative models to create artificial images (Figure 1) that could be used to balance the training dataset.

Although the overall precision of the model didn't improve significantly, styles that were previously underrepresented saw an improvement in classification precision (Figure 2). Conversely, classes with ample real data experienced a reduction in precision. This phenomenon suggests that while balancing the dataset helped underrepresented styles, it may have introduced a bias against styles with abundant data. This differential impact underscores the potential of selectively incorporating artificial data in future work to optimize model performance and enhance architectural style identification.

Figure 1: Generated Images

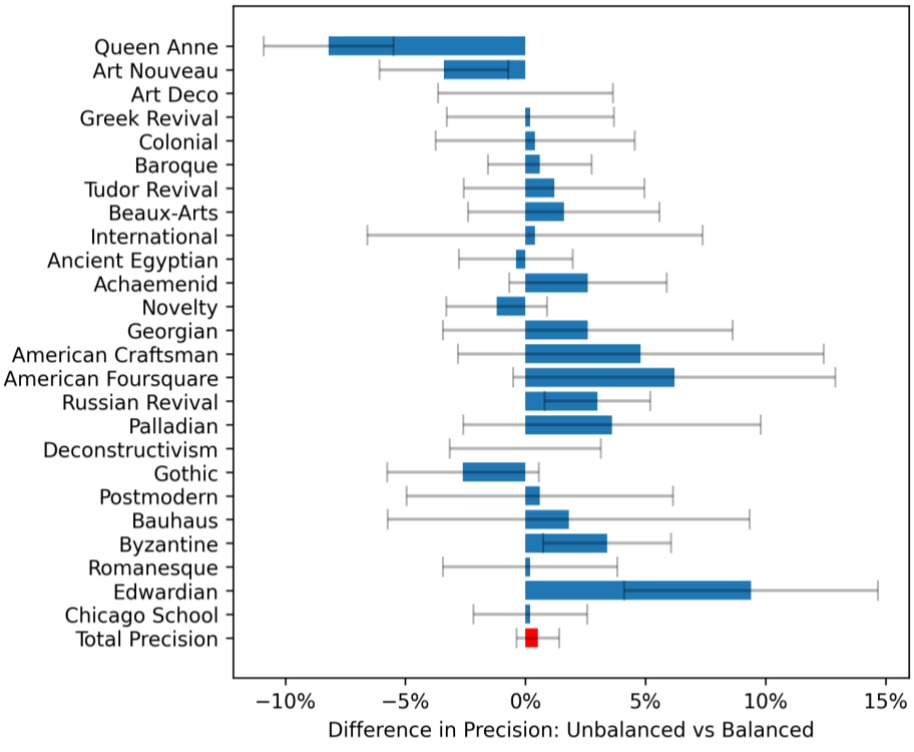

Figure 2: Changes in Precision between Balanced vs Unbalanced Dataset (ordered from the largest class to the smallest class)