Supplementary Material

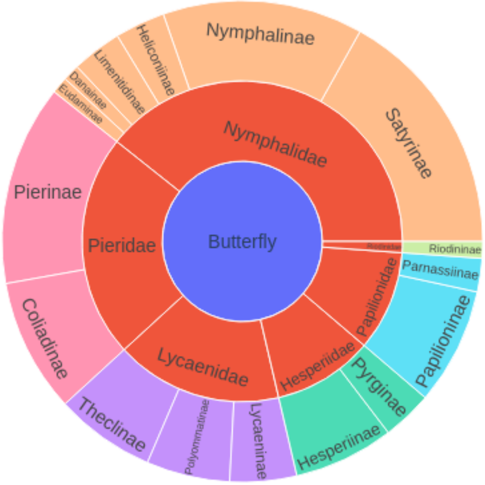


Figure 15. **Butterfly family taxonomy of Fineview dataset**.

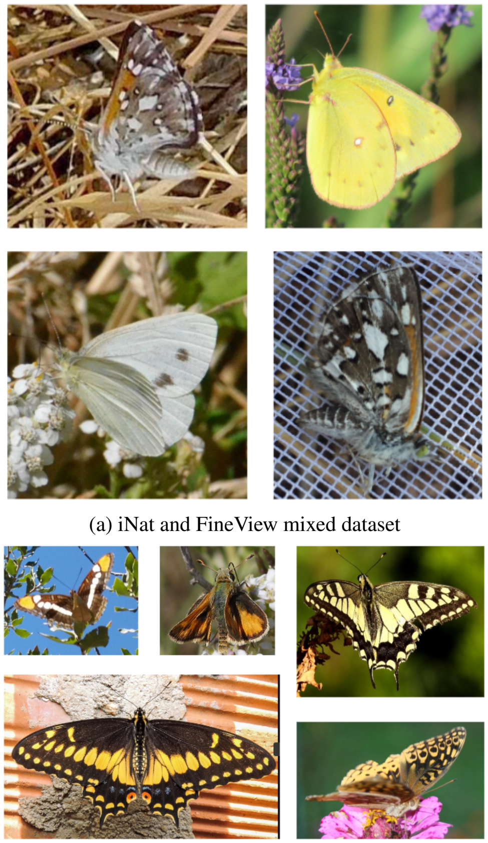
A. FineView dataset taxonomy

In taxonomy, the family rank in the classification of organisms is between genus and order, which is grouped by their common attributes. Butterflies in the same family have some common features, such as shape and color, and it would be subsidiary information for FGVC task. Figure 15 shows the butterfly family and subfamily taxonomy.

B. Further investigation of FGVC task

We investigate the breakdown of incorrect classification of each trained model. Figure 16 shows examples of miss-classified test images of iNat and Fineview mixed dataset-trained model and iNat-only dataset-trained model. The typical misclassified examples of the iNat-only dataset-trained model are certain butterfly poses that extend their wings. This is similar to the butterfly pose of the FineView dataset. The major misclassified examples by The mixed dataset-trained model are the self-occluded butterfly (only certain sides are visible) and the closed-wing pose butterflies.

These results indicate the mixed dataset-trained model accuracy is better than the iNat-only model for certain object pose cases because adding the Fineview dataset reinforces the variety of pose distribution of the training dataset when we use a simple Resnet classification model, and this supports the hypothesis that the classification accuracy depends on the object pose distribution of the training dataset. Furthermore, the FineView mixed-trained model is better especially when the base training dataset is scarce, these results suggest a well-distributed camera pose of the training dataset is crucial for the FGVC task. The FineView dataset is effective for FGVC tasks although the butterfly of the FineView dataset lacks object pose variation. For future work, the FineView dataset can be applied to the object pose-aware classification models for FGVC tasks, which could potentially improve classification accuracy.



(a) iNat and FineView mixed dataset

(b) iNat-only dataset

Figure 16. **Examples of incorrect classification of each trained model**

C. Additional Nerf model examples

One of the advantages of the FineView dataset is the sphere angle distribution of captured images, which is bidirectional 360-degree camera poses. Figure 17 shows several unseen views of synthetic Nerf model-generated images. This camera pose trajectory is along one direction from top to bottom of the sphere of a butterfly. The left column images (*from top image to bottom*) are from top to front view angle and the right column images are from front to bottom view angle camera poses. A particularly eye-catching result is that the butterfly object is invisible in the front view angle camera pose image (*the right top im-*

age in Figure 17). We show the comparison of unseen views of Nerf-generated images (*even rows*) and ground truth test images (*odd rows*) in Figure 18. Horizontal view images (*center column*) are relatively unclear compared to other view images visually and PSNR, SSIM, and LPIPS are approximately 5% worse than other views. We assume the vanilla Nerf model can not capture the horizontal view of the butterfly's body because the butterfly has thin and flat shapes and the antennas and legs are invisible in all generated images. Those flat shapes and fine structures are challenging not only for Nerf models but also for general 3D reconstruction and 3D modeling tasks, and it is a significant research topic for the computer vision community. This is another potential use case of the FineView dataset.

D. FineView dataset examples

Figure 19 shows several sets of examples of multi-view 2D RGB, mask, and the corresponding images. These images have the same pinholder location but are captured by 8 cameras. The mask images capture small structures of butterflies, such as antennae and wing shape. The corresponding key points are consistent between different views. This auxiliary information is labor-intensive for human annotators, but Our proposed system can automatically capture those images.

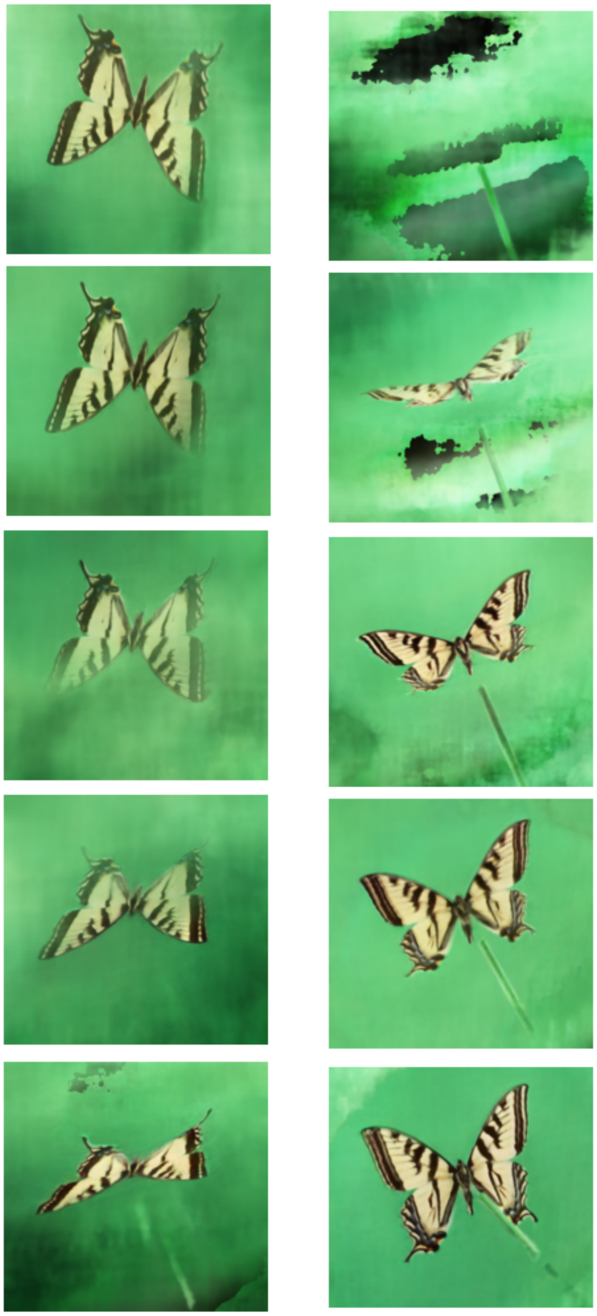


Figure 17. **Various unseen views of Nerf generated images**.

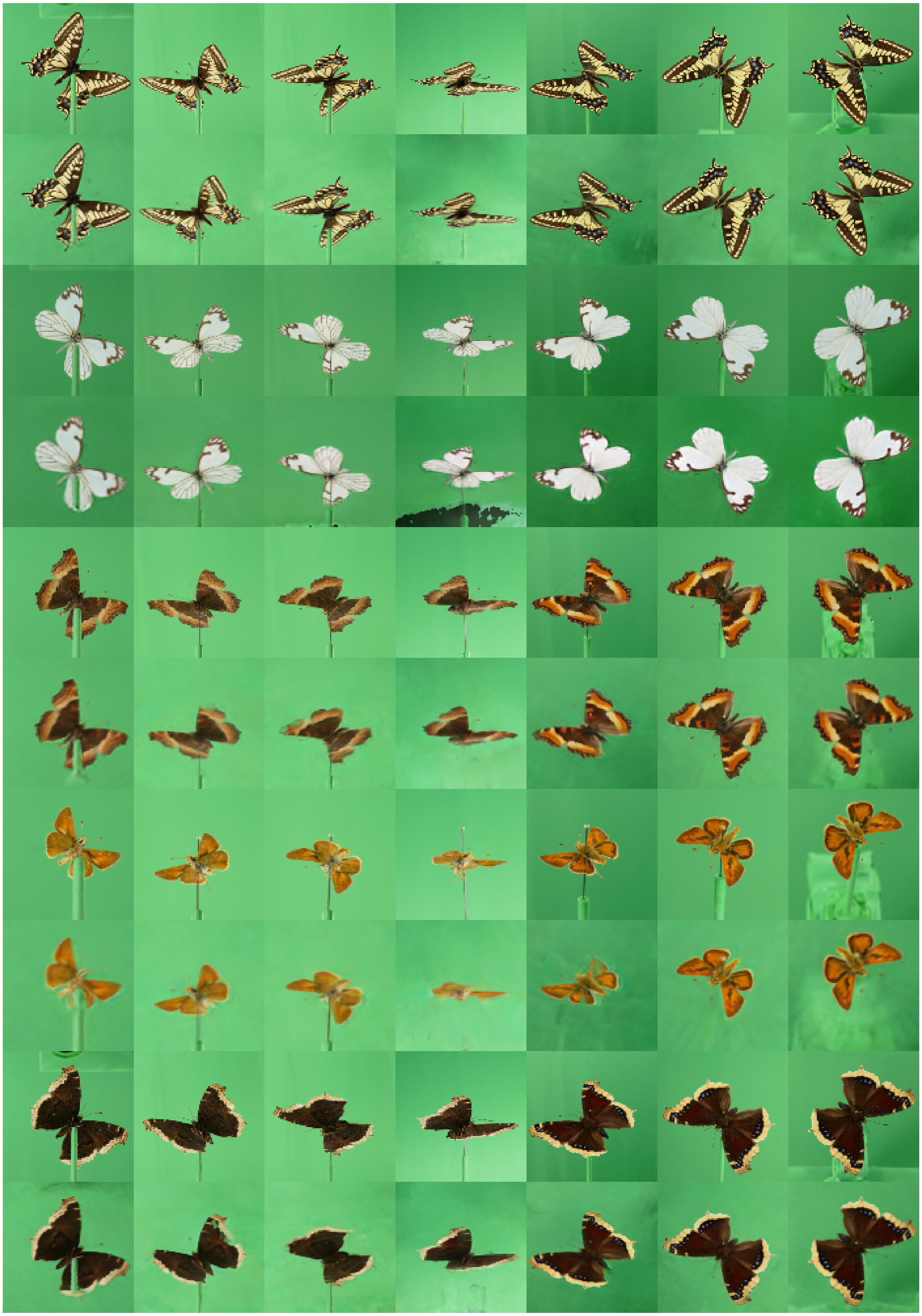


Figure 18. **Ground Truth images (*odd rows*) vs untrained view of Nerf generated images (*even rows*).**

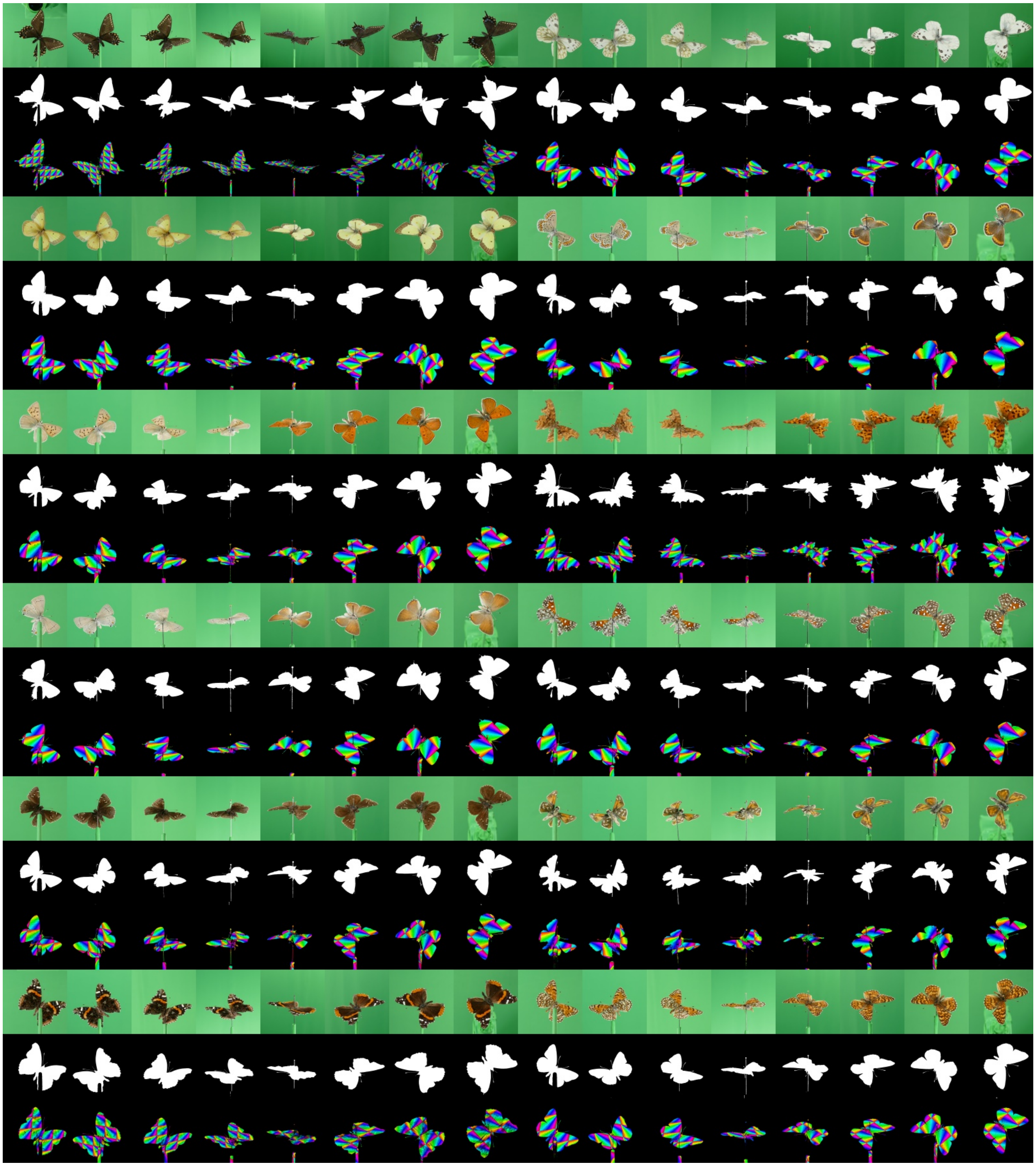


Figure 19. **Examples of multi-view 2D RGB, mask and corresponding images**.