In this supplementary material, we provide our full analysis on AWA, SUN, CUB and FLO datasets. We start our analysis with reporting ZSL and GZSL results on all these datasets evaluating different aspects such as the feature generation model, the classification model, stability, increasing the number of generated features. Moreover, we provide the generated images with StackGAN [6] on CUB and FLO datasets to provide a means for manual inspection.

![Figure 1](image1.png) **Figure 1:** Comparing f-xGAN versions with f-GMMN as well as comparing classification methods in ZSL

![Figure 2](image2.png) **Figure 2:** Comparing f-xGAN versions with f-GMMN as well as comparing classification methods in GZSL

**ZSL and GZSL with f-xGAN.** In this section, we show that our conclusions in the main paper hold across all the datasets by comparing different generative models, i.e. f-xGAN and f-GMMN, and different classification methods, i.e. state of the art multimodal embedding models and softmax, in both ZSL (Figure 1) and GZSL (Figure 2) settings. The full numerical results are presented in Table 1.

First, we compare 'none', i.e. the baseline with no generated features, with different feature generating models. Our observation is that generated features improve over 'none' in both ZSL and GZSL for all embedding methods and all generative models on all the datasets. This indicates that generating features for unseen classes is not only an effective way to balance seen and unseen classes performance in GZSL, but also can boost the ZSL performance.

Second, comparing different generative models, our proposed model achieves the best results in most of the cases, which demonstrates the effectiveness of our novel feature generating f-CLSWGAN model. These results are also worthwhile as they demonstrate the robustness of our feature generation across datasets with varying sizes and granularity.

Third, comparing different classification models, we find that multimodal embedding methods outperform multimodal embedding methods in GZSL. Although 'none' cannot be applied to either ZSL or GZSL settings with softmax classifier, with feature generation softmax is applicable. Even though it is a simple classifier, we show significant improvements over the state of the art with softmax in GZSL setting.

**Seen accuracy vs. training epochs.** In the main paper, we analyze how well different generative models fit the seen class data used for training on CUB and FLO. Here we present our full results on CUB, FLO, AWA and SUN in Figure 3. As we analyzed CUB and FLO in the main paper, there we focus on SUN and AWA. On SUN, we observe that both f-CLSWGAN and f-GMMN almost reach the same performance as the real data. f-WGAN is slightly below them and f-GAN is weaker with a big gap compared to other generative models, which indicates that f-GAN fails to capture the distribution of the training data, i.e. CNN features of seen classes. On AWA, the figure shows that all the generative models are able to fit the distribution of training data although f-GAN converges slower to a slightly worse solution. These experiments demonstrate that our f-CLSWGAN generates robust features that capture the data distribution of all the datasets.

**Unseen accuracy vs. number of generated features.** Here we evaluate the generalization ability of the generative models,
Table 1: ZSL measuring per-class average Top-1 accuracy (T1) on $\gamma^u$ and GZSL measuring $u = T1$ on $\gamma^u$, $H = \text{harmonic mean}$ (FG=feature generator, none: no access to generated CNN features, hence softmax is not applicable). $f$-CLSWGAN significantly boosts both the ZSL and GZSL accuracy of all classification models on all four datasets.

$f$-xGAN and $f$-GMMN, to generate unseen class features for ZSL in Figure 4. On all datasets, $f$-CLSWGAN outperforms other generative models, which demonstrates that our proposed $f$-CLSWGAN is able to generate more robust CNN features for unseen classes. Although $f$-GMMN obtains similar results as $f$-CLSWGAN on CUB and SUN, it is obviously worse than $f$-CLSWGAN on FLO and AWA. Interestingly, we observe that $f$-GAN performs quite well on AWA, which can be explained by its good ability to fit the training data as it is also demonstrated in Figure 3. We conclude from these experiments that our $f$-CLSWGAN generalizes better to unseen classes than other generative models.

**Accuracy vs. number of generated features.** In Figure 5 we analyze how the number of generated features affects results measured in harmonic mean accuracy in GZSL. We train a softmax classifier with real features of seen classes and generated features unseen classes, and evaluate its performance by calculating the harmonic mean of seen and unseen class accuracies. First of all, compared to the ZSL setting in Figure 4 we find that more generated features are needed to achieve the optimal
harmonic mean, which is reasonable because GZSL suffers from data imbalance issues and more generated unseen class features are required to balance seen and unseen class accuracies. Second, our \texttt{f-CLSWGAN} attains the best results on all the datasets and the ranking is roughly \texttt{f-CLSWGAN} > \texttt{f-GMMN} > \texttt{f-WGAN} > \texttt{f-GAN}. Again, we conclude that our \texttt{f-CLSWGAN} is more effective to balance seen and unseen class accuracies in GZSL than other generative models.

**Visualization of generated 256 \times 256 images.** In the main paper we provide quantitative results and conclude that extracting CNN features from generated images instead of directly generating CNN features fails to generalize to unseen classes. Here, we analyze this conclusion via manual inspection of the generated images.

Figure 7 shows samples of randomly chosen images from 20 classes of the CUB dataset. While most images are recognizable as birds, their type varies and does not resemble a consistent class. Only few samples share similar properties with birds from real images (left column). Generated images from the FLO dataset (Figure 6) often have similar colors as the original. However other properties such as shape, size and count of petals vary significantly and differ from the original class.

We conjecture that the inconsistencies within a class, as well as the mismatch between generated and real images are causes why training on generated images gives poor results.

**References**


Figure 6: Visualization of generated $256 \times 256$ images by StackGAN [6] on FLO. Although generated images share similar colors with real ones, most of them fail to resemble correct shapes, which leads to inferior classification results.
Figure 7: Visualization of generated $256 \times 256$ images by StackGAN [6] on CUB. Although generated images can be recognized as birds, many of them fail to capture structural details and are class-inconsistent.