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.

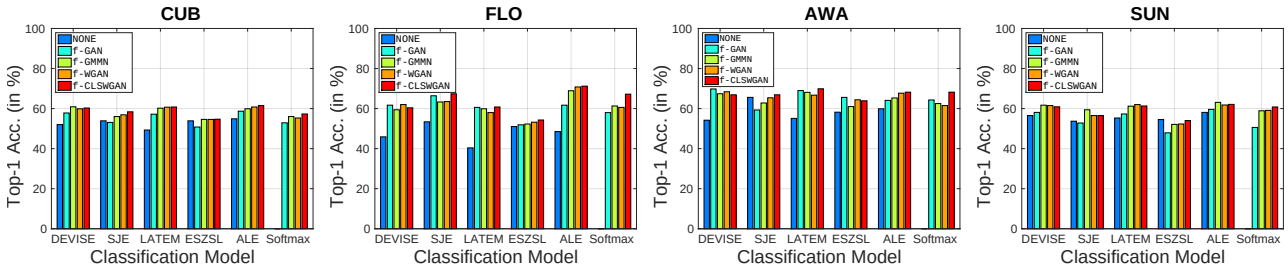

Figure 1: Comparing f−xGAN versions with f−GMMN as well as comparing classification methods in ZSL

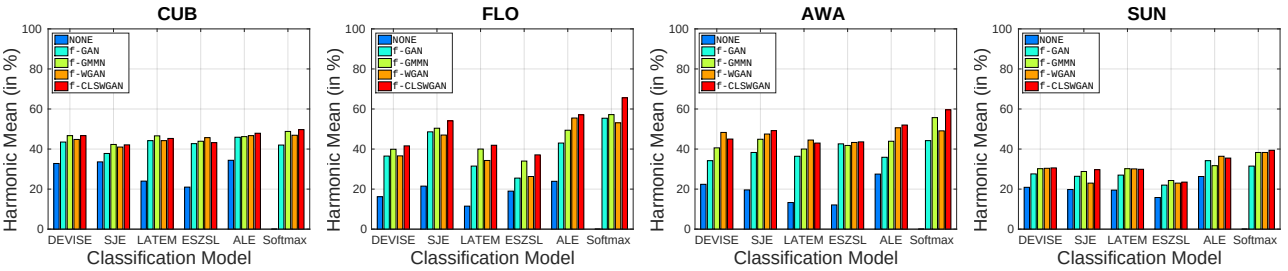

Figure 2: Comparing f−xGAN versions with f−GMMN as well as comparing classification methods in GZSL

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 obtain better results in ZSL and softmax classifier outperforms multimodal embedding methods in GZSL. Although 'none' can not 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,

		Zero-Shot Learning				Generalized Zero-Shot Learning											
		CUB	FLO	SUN	AWA	CUB			FLO			SUN			AWA		
Classifier	FG	T1	T1	T1	T1	u	s	H	u	s	H	u	s	H	u	s	H
DEVISE [3]	none	52.0	45.9	56.5	54.2	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9	13.4	68.7	22.4
	f-GAN	57.8	61.7	58.1	69.8	45.2	41.9	43.5	44.8	30.8	36.5	36.6	22.2	27.6	22.0	76.8	34.2
	f-GMMN	60.9	59.4	61.7	67.4	51.1	43.0	46.7	44.8	35.9	39.9	36.1	26.0	30.2	27.6	76.1	40.6
	f-WGAN	59.9	62.0	61.5	68.4	51.8	39.4	44.8	48.9	29.2	36.6	42.8	23.6	30.4	38.9	63.8	48.3
	f-CLSWGAN	60.3	60.4	60.9	66.9	52.2	42.4	46.7	45.0	38.6	41.6	38.4	25.4	30.6	35.0	62.8	45.0
SJE [2]	none	53.9	53.4	53.7	65.6	23.5	59.2	33.6	13.9	47.6	21.5	14.7	30.5	19.8	11.3	74.6	19.6
	f-GAN	53.1	66.4	52.8	59.3	46.0	32.1	37.8	55.9	42.9	48.6	39.2	19.9	26.4	30.5	51.6	38.3
	f-GMMN	56.0	63.3	59.4	62.8	49.3	37.1	42.3	52.4	48.5	50.4	47.2	20.8	28.8	32.8	71.1	44.9
	f-WGAN	56.9	63.5	56.5	65.4	48.6	35.4	41.0	55.4	40.8	47.0	19.8	27.7	23.1	35.9	70.3	47.5
	f-CLSWGAN	58.4	67.4	56.5	66.9	48.1	37.4	42.1	52.1	56.2	54.1	36.7	25.0	29.7	37.9	70.1	49.2
LATEM [5]	none	49.3	40.4	55.3	55.1	15.2	57.3	24.0	6.6	47.6	11.5	14.7	28.8	19.5	7.3	71.7	13.3
	f-GAN	57.2	60.6	57.3	69.0	45.4	43.0	44.2	49.0	23.2	31.5	35.9	21.6	27.0	23.9	75.9	36.4
	f-GMMN	60.2	59.9	61.2	68.1	52.3	42.0	46.6	44.9	36.0	40.0	35.7	26.2	30.2	27.2	76.0	40.0
	f-WGAN	60.7	58.0	62.0	66.7	54.6	37.1	44.2	50.2	26.0	34.3	42.6	23.3	30.1	34.6	62.4	44.5
	f-CLSWGAN	60.8	60.8	61.3	**69.9**	53.6	39.2	45.3	47.2	37.7	41.9	42.4	23.1	29.9	33.0	61.5	43.0
ESZSL [4]	none	53.9	51.0	54.5	58.2	12.6	63.8	21.0	11.4	56.8	19.0	11.0	27.9	15.8	6.6	75.6	12.1
	f-GAN	50.8	51.9	47.9	65.6	38.9	47.3	42.7	16.8	53.1	25.5	31.8	16.9	22.0	29.8	74.7	42.6
	f-GMMN	54.6	52.3	52.1	61.0	48.3	40.2	43.9	22.6	67.9	34.0	29.4	20.7	24.3	28.9	75.4	41.8
	f-WGAN	54.6	53.2	52.3	64.4	45.6	45.7	45.7	17.5	53.2	26.3	41.5	15.9	23.0	30.6	74.3	43.3
	f-CLSWGAN	54.7	54.3	54.0	63.9	36.8	50.9	43.2	25.3	69.2	37.1	27.8	20.4	23.5	31.1	72.8	43.6
ALE [1]	none	54.9	48.5	58.1	59.9	23.7	62.8	34.4	13.3	61.6	21.9	21.8	33.1	26.3	16.8	76.1	27.5
	f-GAN	58.7	61.7	59.6	64.1	46.0	45.9	45.9	54.8	35.4	43.0	40.3	29.7	34.2	24.9	64.5	35.9
	f-GMMN	59.9	68.9	63.1	65.3	48.9	44.9	46.2	40.2	64.1	49.4	28.7	35.5	31.7	33.9	62.4	43.9
	f-WGAN	60.8	70.8	61.8	67.7	51.9	42.4	46.7	57.0	54.1	55.5	44.2	30.9	36.4	43.6	60.3	50.6
	f-CLSWGAN	**61.5**	**71.2**	**62.1**	68.2	40.2	59.3	47.9	54.3	60.3	57.1	41.3	31.1	35.5	47.6	57.2	52.0
Softmax	none	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	f-GAN	52.9	58.0	50.6	64.3	36.0	50.3	42.0	45.7	70.3	55.4	34.3	29.1	31.5	36.0	57.4	44.2
	f-GMMN	56.0	61.3	58.9	62.5	43.4	55.9	48.8	47.8	71.0	57.2	36.0	41.0	38.3	48.2	66.1	55.7
	f-WGAN	55.3	60.6	59.1	61.5	51.4	43.2	46.9	43.4	68.5	53.1	42.6	34.8	38.3	43.2	56.8	49.1
	f-CLSWGAN	57.3	67.2	60.8	68.2	43.7	57.7	**49.7**	59.0	73.8	**65.6**	42.6	36.6	**39.4**	57.9	61.4	**59.6**

Table 1: ZSL measuring per-class average Top-1 accuracy (T1) on $\mathcal{Y}^u$ and GZSL measuring $\mathbf{u}$ = T1 on $\mathcal{Y}^u$, $\mathbf{s}$ = T1 on $\mathcal{Y}^s$, H = 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

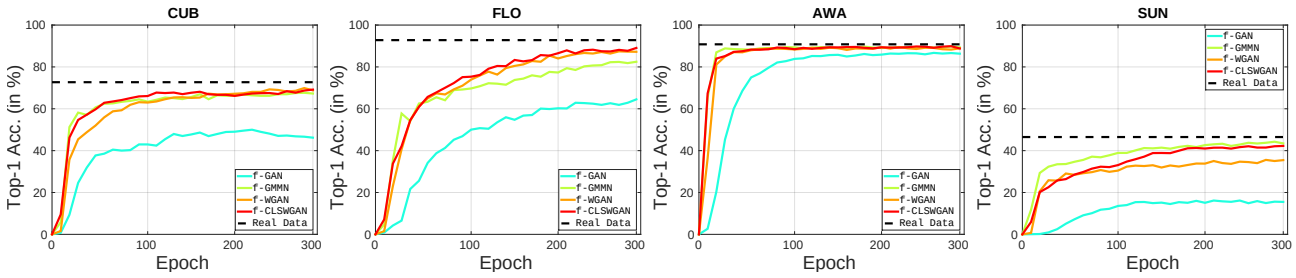

Figure 3: Seen class accuracy of the softmax classifier trained on generated features w.r.t. the training epochs.

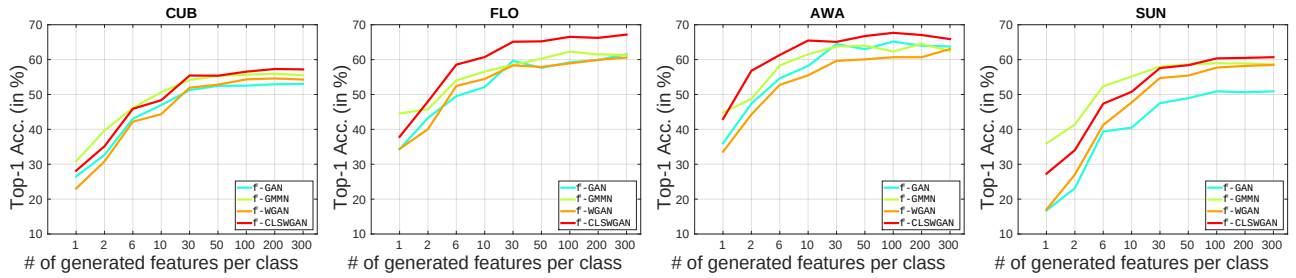

Figure 4: Increasing the number of generated `f-xGAN` features w.r.t. unseen class accuracy (with `softmax`) in ZSL.

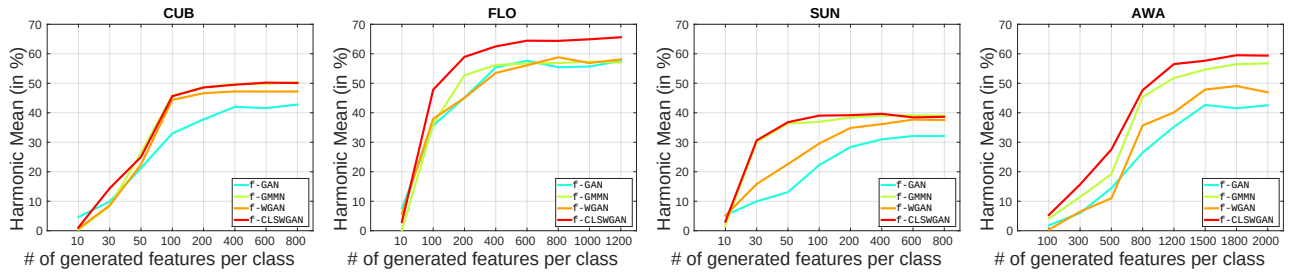

Figure 5: Increasing the number of generated `f-xGAN` features w.r.t. harmonic mean (with `softmax`) in GZSL

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 `f-CLSWGAN` attains the best results on all the datasets and the ranking is roughly `f-CLSWGAN` $>$ `f-GMMN` $>$ `f-WGAN` $>$ `f-GAN`. Again, we conclude that our `f-CLSWGAN` is more effective to balance seen and unseen class accuracies in GZSL than other generative models.

Visualization of generated 256×256 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

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 2016.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[3] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[4] B. Romera-Paredes and P. H. Torr. An embarrassingly simple approach to zero-shot learning. *ICML*, 2015.

[5] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[6] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

Figure 6: Visualization of generated 256×256 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 256×256 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.