Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification (Supplementary)

Sanath Narayan*[1], Akshita Gupta*[1], Fahad Shahbaz Khan[1,3], Cees G.M. Snoek[2], Ling Shao[1,3]

[1] Inception Institute of Artificial Intelligence, UAE [2] University of Amsterdam
[3] Mohamed Bin Zayed University of Artificial Intelligence, UAE

In this supplementary material, we present additional quantitative and qualitative analysis of the zero-shot recognition performance of our `TF-VAEGAN` framework. While the additional quantitative results are presented in Sec. 1, the qualitative results are discussed in Sec 2 of this supplementary.

1 Quantitative Results

In this section, we present the ablation studies with respect to the feedback design choices and the choice of latent embeddings.

Feedback design choices: Here, we explore the effect of changing the input to the feedback module F and its associated training strategy on CUB. Originally, the input to F is taken from discriminator D and the training of F is performed in a two-stage strategy. This setup is denoted by `TwoStage+D` and obtains classification performance of 61.4% and 53.3% for ZSL and GZSL. Instead, in our approach, the input to F is taken from SED *Dec*. This setup is denoted by `TwoStage+Dec` and achieves performance of 62.0% and 53.8% for ZSL and GZSL. Further, we utilize an alternate training strategy combined with `TwoStage+Dec` to facilitate the generator training, thereby improving feature synthesis. This setup, denoted by `Our Feedback`, achieves improved performance of 62.8% and 54.8% for ZSL and GZSL. These results show that (i) `TwoStage+Dec` provides improved performance over original `TwoStage+D` and (ii) the best results are obtained by `Our Feedback`, demonstrating the impact of our modifications for improved zero-shot recognition.

Choice of latent embeddings for `T-feature`: Here, we evaluate the impact of concatenating different embeddings from SED to the baseline features. We compare our proposed concatenation (`T-feature`) of baseline features with latent embeddings h of SED with both the original baseline features (`OrigFeat`) and the baseline features concatenated with the reconstructed attributes (`ConcatFeat`). On CUB, `OrigFeat` achieves 61.2% and 53.5% on ZSL and GZSL tasks, respectively. `ConcatFeat` achieves gains of 1.6% and 2.0% over `OrigFeat`. In case of `ConcatFeat`, the reconstructed attributes have single feature representations per-class with inter-class separability but no intra-class diversity. Different to reconstructed attributes, the latent embeddings h possess both intra-class diversity (multiple feature instances per class) and inter-class separability. Our

`T-feature` exploits these properties of latent embeddings with improved results over both `OrigFeat` and `ConcatFeat`. Compared to `OrigFeat`, `T-feature` obtains gains of 2.8% and 3.4% on ZSL and GZSL tasks, respectively.

2 Qualitative Analysis

2.1 Feature Visualization Comparison

Here, we present the implementation details and additional qualitative results for the visualization of synthesized features discussed in Sec. 4.2 of the paper.
Implementation details: The image generator, which inverts the feature instances to images of size 64x64, consists of a fully-connected (FC) layer followed by five upconvolutional blocks. Each upconvolutional block contains an Upsampling layer, a 3x3 convolution, BatchNorm and ReLu non-linearity. An ℓ_1 loss between the ground truth and inverted images, along with a perceptual loss (ℓ_2 loss between the corresponding feature vectors at conv5 of a pre-trained ResNet-101) and an adversarial loss are employed to construct good quality images. The discriminator, required for adversarial training, takes image and feature embedding as inputs. The input image is processed through four downsampling blocks to obtain an image embedding, while the feature embedding is passed through an FC layer and spatially replicated to match the spatial dimensions of the obtained image embedding. The resulting two embeddings are concatenated and passed through convolutional and *sigmoid* layers for predicting whether the input image is real or fake. The model is trained on all the real feature-image pairs of the 102 classes of FLO [1].
Visualization: The comparison between `Baseline` and our `Feedback` synthesized features on eight example flowers is shown in Fig. 1. For each flower class, a ground-truth (GT) image along with three images inverted from its GT feature, `Baseline` and `Feedback` synthesized features, respectively are shown. Generally, inverting the `Feedback` synthesized feature yields an image that is semantically closer to the GT image than inverting the `Baseline` synthesized feature. Inverting the feature instances from our `Feedback` improves the color of bud and shape of petals (*Californian poppy*, *Globe flower* and *Osteospermum*), structure of the flower (*Hippeastrum*), in comparison to the `Baseline` synthesized features. A considerable improvement for our `Feedback` over the `Baseline` is visible in these flowers (*Californian poppy*, *Globe flower*, *Hippeastrum* and *Osteospermum*). However, there are a few challenging cases (*e.g.*, *Globe thistle*, *Windflower*, *Sweet william*, *Moon orchid*), where a semantic gap still exists between the inversion of real features (denoted as `Reconstructed`) and inversion of `Feedback` synthesized features, even though there is a marginal improvement for our `Feedback` over the `Baseline`. These qualitative observations suggest that our `Feedback` improves the feature synthesis stage over the `Baseline`, where no feedback is present, resulting in improved zero-shot classification.

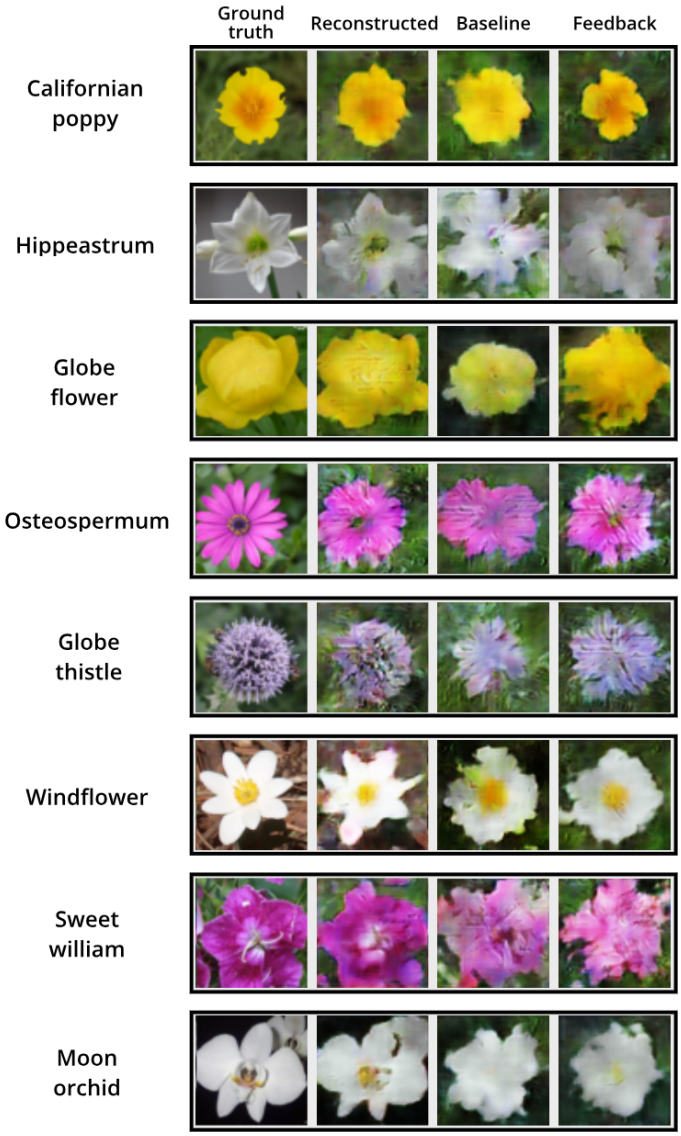

Fig. 1: Qualitative comparison between inverted images of `Baseline` synthesized features and our `Feedback` synthesized features on eight example classes of FLO [1]. The ground-truth image (GT) and the reconstructed inversion (`Reconstructed`) of its real feature are also shown for each example. Inverting the feature instances from our `Feedback` improves the color of bud and shape of petals (*Californian poppy*, *Globe flower* and *Osteospermum*), structure of the flower (*Hippeastrum*), in comparison to the `Baseline` synthesized features. Semantic gap still exists between the inversion of real features (denoted as `Reconstructed`) and inversion of `Feedback` synthesized features for a few challenging cases (*e.g.*, *Globe thistle*, *Windflower*, *Sweet william*, *Moon orchid*), even though there is some improvement for our `Feedback` over the `Baseline`. These observations suggest that our `Feedback` improves the quality of synthesized features over the `Baseline`, where no feedback is present. Best viewed in color and zoom.

2.2 Classification Performance Comparison

Here, we qualitatively illustrate the performance of our `TF-VAEGAN` framework, in comparison to the baseline `f-VAEGAN` [3] method, on two fine-grained object recognition datasets: CUB and FLO. Fig. 2 and 3 present the comparison on CUB and FLO, respectively. For each dataset, images from five most confusing categories (with respect to the baseline `f-VAEGAN`) are shown. The comparison is illustrated for five image instances in each category. The ground truth instances are shown in the top row for each category, followed by the classification results of the baseline and proposed frameworks in second and third rows, respectively. Correctly classified images are marked with a green border, while the incorrectly classified images are marked with a red border. For the misclassifications, the name of the incorrectly predicted class is denoted below the instance for the respective methods.

CUB: The qualitative comparison between the baseline and the proposed approaches for the CUB [2] dataset is shown in Fig. 2. Five categories of birds that are most confusing for the baseline approach are presented. The categories are *Prairie warbler*, *Great crested flycatcher*, *Grovve billed ani*, *Herring gull* and *Califonria gull*. Generally, for all these categories, the baseline `f-VAEGAN` approach confuses with similar looking bird categories in the dataset. Our `TF-VAEGAN` reduces this confusion between similar looking classes and improves the classification performance. In Fig. 2, we observe that the baseline approach confuses *Prairie warbler* class with other similar looking *warbler* categories such as *Blue winged warbler*, *Magnolia warbler* and *Orange crowned warbler*. This confusion is reduced in the predictions of our `TF-VAEGAN`. Similarly, the confusion present, in the baseline method, between the *Great crested flycatcher* and other *flycatcher* categories is reduced for the proposed method. As a result, the overall classification performance improves for the proposed method over the baseline.

FLO: Fig. 3 shows the qualitative comparison for five categories of flowers from the Oxford Flowers [1] dataset that are most confusing for the baseline method. The categories are *Dafodil*, *Pink primrose*, *Siam tulip*, *King Protea* and *Common dandelion*. For all these categories, the proposed `TF-VAEGAN` reduces the confusion present between the similar looking classes in the baseline `f-VAEGAN` approach and improves the classification performance. In general, we observe that the instances are misclassified to other similar looking categories in the dataset. *E.g.*, instances of *Common dandelion* are commonly misclassified as either *Colt's foot* or *Yellow iris*. All three categories have yellow flowers and share similar appearance. We observe that the baseline makes confused predictions with respect to these classes. However, the confusion is less in the predictions of the proposed `TF-VAEGAN`. This leads to a favourable improvement in the zero-shot classification performance for the proposed approach. Similar observations can also be made in the case of other categories. The baseline `f-VAEGAN` generally confuses *Dafodil* with *Globe flower* and *Yellow iris* due to the yellow colour, while *Pink primrose* is mostly confused with *Petunia* and *Monkshood* due to the pinkish petals in the flowers. The misclassifications are reduced when using the proposed `TF-VAEGAN` for classification, resulting in an improved performance.

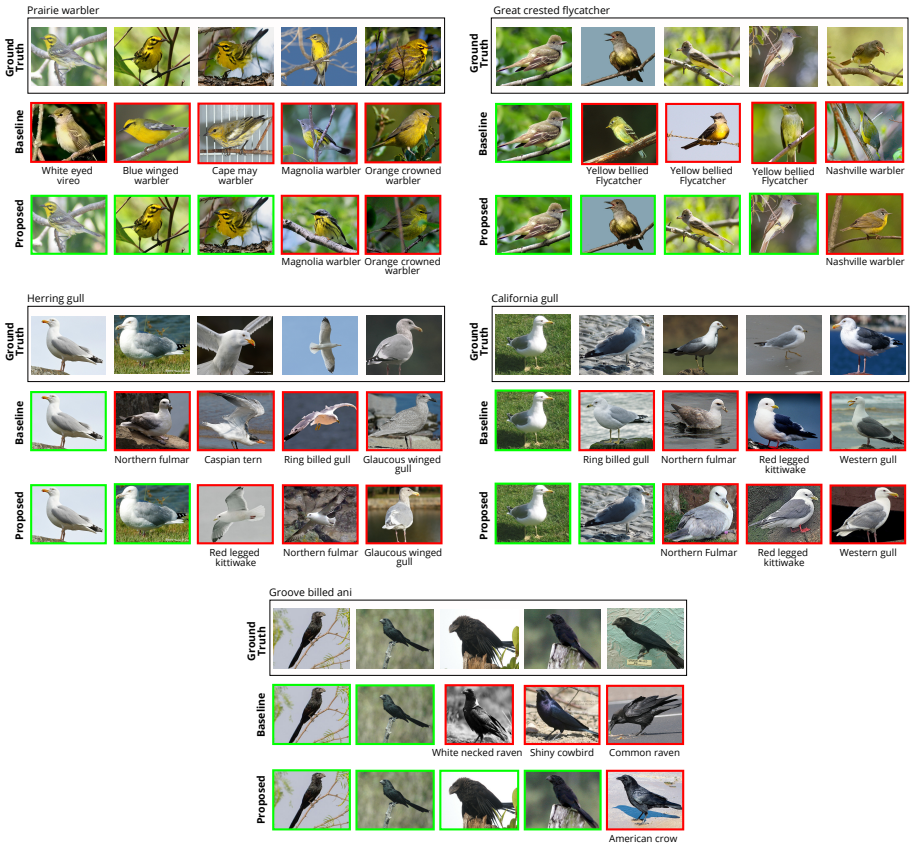

Fig. 2: Qualitative comparison between the baseline and our proposed approach on the CUB [2] dataset. The comparison is based on the most confusing categories as per the baseline performance. For each category, while the top row denotes different variations of ground truth class instances, the second and third rows show the classification predictions by the baseline and proposed approaches, respectively. The green and red boxes denote correct and incorrect classification predictions, respectively. The class names under each red box show the corresponding incorrectly predicted label. In general, we observe that the instances are misclassified to other similar looking categories in the dataset. For instance, *Prairie warbler* is confused with *Blue winged warbler*, while *Groove billed ani* is confused commonly with *Common raven*. For all these categories, the proposed `TF-VAEGAN` reduces the confusion among similar looking classes in the baseline `f-VAEGAN` and improves the classification performance over the baseline. See associated text for additional details. Best viewed in color and zoom.

Fig. 3: Qualitative comparison between the baseline and our proposed approach on the Oxford Flowers [1] dataset. The comparison is based on the most confusing categories as per the baseline performance. For each category, while the top row denotes different variations of ground truth class instances, the second and third rows show the classification predictions by the baseline and proposed approaches, respectively. The green and red boxes denote correct and incorrect classification predictions, respectively. The class names under each red box show the corresponding incorrectly predicted label. In general, we observe that the instances are misclassified to other similar looking categories in the dataset. For instance, *Common dandelion* is confused with *Colt's foot*, while *Pink primrose* is confused with *Petunia*. For all these categories, the proposed `TF-VAEGAN` reduces the confusion among similar looking classes in the baseline `f-VAEGAN` and improves the classification performance over the baseline. See associated text for additional details. Best viewed in color and zoom.

References

1. Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2, 3, 4, 6
2. Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech*, 2010. 4, 5
3. Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 4