

A Novel Baseline for Zero-shot Learning via Adversarial Visual-Semantic Embedding (supplementary material)

Yu Liu

yu.liu@esat.kuleuven.be

Tinne Tuytelaars

tinne.tuytelaars@esat.kuleuven.be

ESAT-PSI

KU Leuven

Leuven, Belgium

1 Network details for AVSE

As shown in Fig.2 in the paper, AVSE is built with several components, including encoder, generator, discriminator, classifier and regressor. We implemented the components with multi-layer perceptrons (MLPs). Table 1 below describes the details in these components. Notice that, each component contains one or two fully-connected layer, thereby it is effective to train the entire model in an end-to-end fashion.

Encoder (E)	Generator (G)	Discriminator (D)	Classifier (C)	Regressor (R)
FC-4096	FC-4096	FC-4096	FC-#classes	FC-#attributes
Leaky ReLU (0.2)	Leaky ReLU (0.2)	Leaky ReLU (0.2)	Softmax	ReLU
FC-2048	FC-2048	FC-2		
ReLU	ReLU			

Table 1: Network details in the AVSE model. FC represents the fully-connected layer, and the number behind it denotes the number of output units. ‘#classes’ is the number of seen classes and ‘#attributes’ is the number of semantic attributes.

2 More embedding-to-image generation examples

In Sec.3.3 of the paper, we train the embedding-to-image generation model and show some examples of generated images on Oxford Flowers dataset. Below, we illustrate more visualization examples for seen classes in Fig. 1 and for unseen classes in Fig. 2, respectively. These examples Additionally show the effectiveness of the embedding learned in AVSE.

3 Confusion matrix between seen and unseen classes

One challenge in ZSL is the misclassification between seen and unseen classes. To this end, we compute a distance score to quantify the distributions of seen and unseen class

prototypes (see Sec.4.3 in the main paper). Recall that, we compute an Euclidean distance score for a pair of a seen class prototype and a unseen class one. Consequently, we obtain a $|\mathcal{Y}^u| \times |\mathcal{Y}^s|$ confusion matrix where $|\mathcal{Y}^u|$ is the number of unseen classes and $|\mathcal{Y}^s|$ is the number of seen classes. Based on the confusion matrix, we calculate the average distance score, as reported in Table 5 of the paper. Here, we aim to additionally exhibit the details in the confusion matrix. In Fig. 3, we compare the matrices for f-CLAWGAN, Lis-GAN and AVSE, respectively.

4 Additional classification results

In this supplementary material, we illustrate more classification examples from our AVSE, in the context of both ZSL and GZSL (Fig. 4 for CUB, Fig. 5 for SUN, Fig. 6 for AWA and Fig. 7 for FLO). Note that we show both success and failure cases on each dataset. These examples qualitatively demonstrate the advantage and weakness of AVSE.

5 On the exploration of hard zero-shot learning

Recall the ratio of seen classes to unseen classes in the datasets (see Table 1 in the paper), we can see that ZSL generally defines more seen classes than unseen classes. However, in real-world scenarios, there will be more unseen classes than seen classes. To this end, we propose a harder yet practical setting by switching the numbers of seen and unseen classes. Specifically, we use fewer seen classes to train the model, but more unseen classes to test it. Table 5 reports the results under the hard ZSL setting. Compared with the results in Table 1 and Table 2 of the main paper, we note that the performance drops largely on the datasets for the three methods. We believe there is still much space to further extend the ZSL research in real-world settings.

Method	CUB	SUN	AWA	FLO
f-CLSWGAN	28.1	14.5	17.5	16.1
Lis-GAN	28.5	14.9	18.5	16.9
AVSE	29.5	15.8	18.2	16.9

Table 2: Results of hard zero-shot learning (HZSL), where the number of unseen classes is more than the number of seen classes. The top-1 accuracy of the three methods are competitive on the datasets.

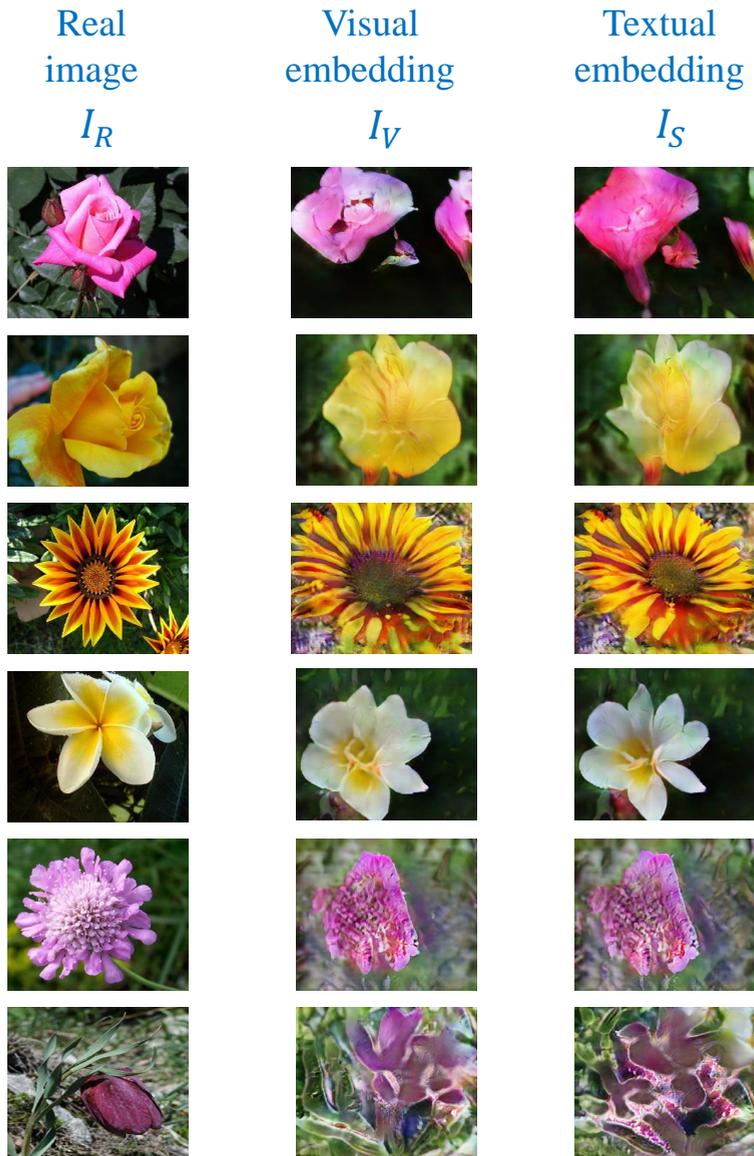


Figure 1: Generated images conditioned on our visual and textual (or semantic) embeddings, respectively. The six images are from **seen classes** of Oxford Flowers. The image instance in the last row looks more challenging than others.

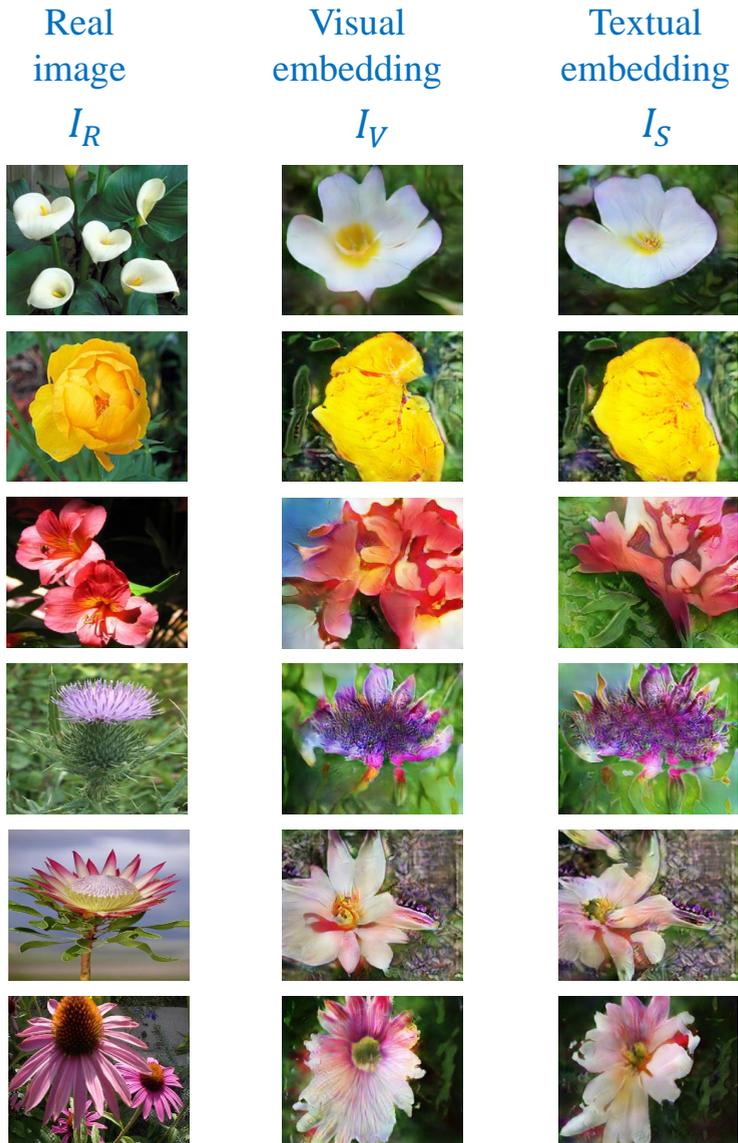


Figure 2: Generated images conditioned on our visual and textual (or semantic) embeddings, respectively. The six images are from **unseen classes** of Oxford Flowers. Compare with the examples in Fig. 1, the image generation for unseen classes is more difficult.

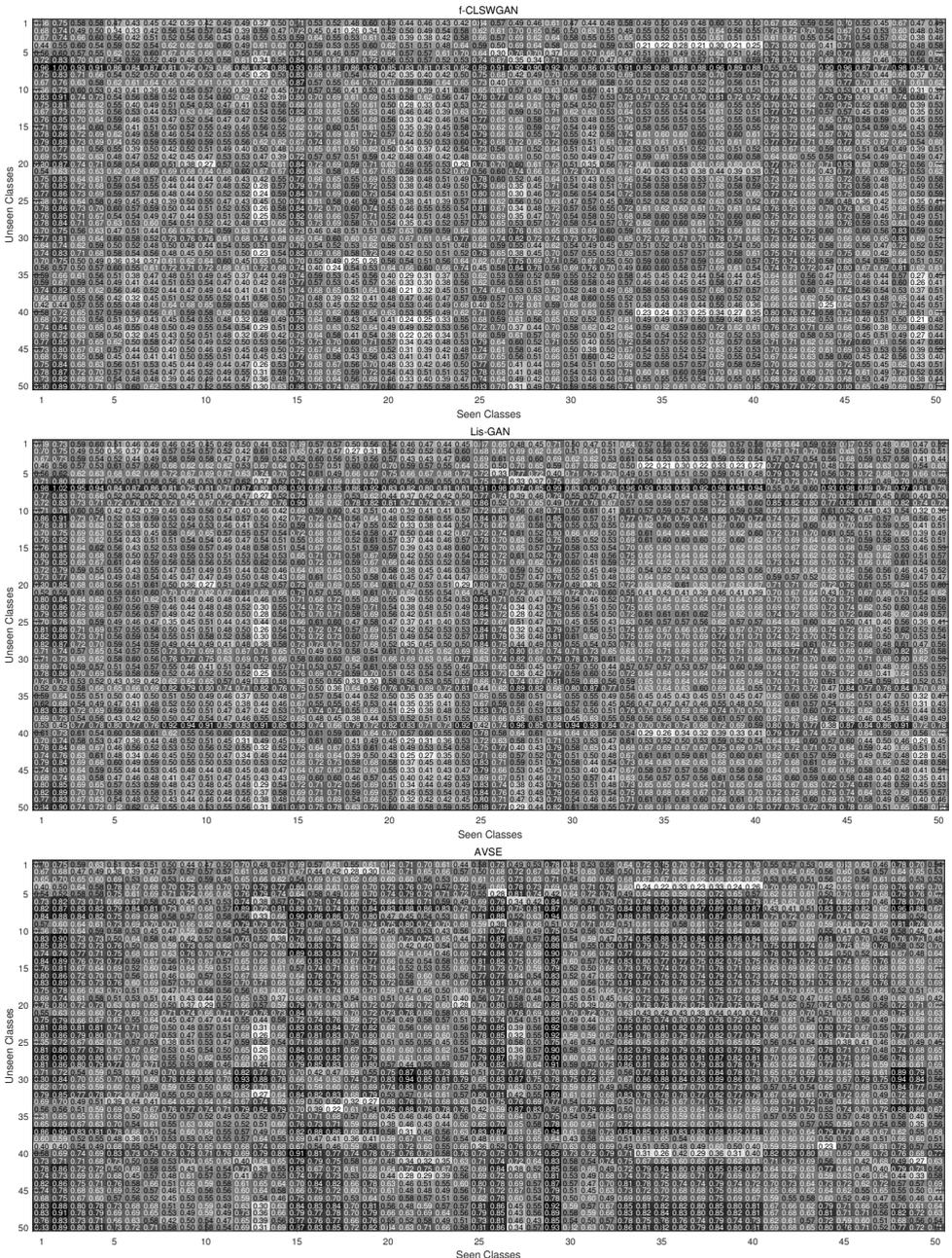


Figure 3: Confusion matrix between unseen class prototypes and seen class prototypes. **Top: f-CLSWGAN; Middle: Lis-GAN; Bottom: AVSE.** We show all the 50 unseen classes, and pick up 50 seen classes from the total 150 seen classes due to the limit of space. (zoom in to read the detailed distance scores in the matrices).

Unseen class	ZSL	GZSL
	<u>Groove billed Ani</u> Bronzed Cowbird Yellow billed Cuckoo Brandt Cormorant Black billed Cuckoo	<u>Groove billed An</u> Fish Crow Shiny Cowbird American Crow Brandt Cormorant
	<u>Black billed Cuckoo</u> Field Sparrow Yellow billed Cuckoo Groove billed Ani Savannah Sparrow	<u>Black billed Cuckoo</u> Field Sparrow Yellow billed Cucko Groove billed Ani Carolina Wren
	<u>Yellow headed Blackbird</u> Groove billed Ani Pileated Woodpecker Red legged Kittiwake Tree Swallow	<u>Yellow headed Blackbird</u> Groove billed Ani Prothonotary Warbler Pileated Woodpecker Red winged Blackbird
	<u>Tropical Kingbird</u> Groove billed Ani Yellow headed Blackbird Yellow bellied Flycatcher Green Violetear	<u>Tropical Kingbird</u> Groove billed Ani Green Violetear Yellow headed Blackbird Yellow bellied Flycatcher
	Red legged Kittiwake Groove billed Ani Yellow billed Cuckoo Caspian Tern Yellow headed Blackbird	Red legged Kittiwake Groove billed Ani Forsters Tern Yellow billed Cuckoo Western Gull

Figure 4: AVSE classification results on **CUB** dataset. In the context of either ZSL or GZSL, the top-5 predictions are shown. Unseen classes are in green and seen classes are in blue. The ground-truth class labels are underlined if the prediction is correct in the top-5. The last image instance belong to the category 'Mallard' is a failure case.

Unseen class	ZSL	GZSL
	<u>vestry</u> geodesic dome indoor mosque indoor pub indoor artists loft	throneroom confessional Seen great hall Seen chapel Seen <u>vestry</u>
	<u>alley</u> observatory outdoor motel rectory ticket booth	ghost town flood observatory outdoor fort tree house
	<u>chemistry lab</u> bank vault exhibition hall ballroom artists loft	airport ticket counter machine shop brewery indoor hangar indoor <u>chemistry lab</u>
	<u>motel</u> rectory yard observatory outdoor playground	hacienda hunting lodge outdoor beach house lawn ranch house
	<u>casino outdoor</u> bog motel pub indoor canal natural	bayou resort pond palace canal natural

Figure 5: AVSE classification results on SUN dataset. In the context of either ZSL or GZSL, the top-5 predictions are shown. Unseen classes are in green and seen classes are in blue. The prediction that is underlined represents the ground-truth class label. Since SUN dataset has more classes than other datasets, its GZSL classification becomes more difficult. It can be seen that seen classes always rank before the unseen classes.

Unseen class	ZSL	GZSL
	<u>seal</u> walrus sheep blue+whale dolphin	otter humpback+whale hippopotamus <u>seal</u> siamese+cat
	<u>horse</u> sheep bobcat blue+whale giraffe	<u>horse</u> cow moose tiger ox
	<u>sheep</u> horse rat walrus bobcat	pig <u>sheep</u> collie cow chihuahua
	blue+whale <u>dolphin</u> seal walrus horse	killer+whale humpback+whale blue+whale <u>dolphin</u> otter
	sheep bat rat horse seal	moose Wolf raccoon otter deer

Figure 6: AVSE classification results on AWA dataset. In the context of either ZSL or GZSL, the top-5 predictions are shown. Unseen classes are in green and seen classes are in blue. The prediction that is underlined represents the ground-truth class label. The last unseen class 'bobcat' fails to be classified as it is too small in the image.

Unseen class	ZSL	GZSL
	<p><u>pink primrose</u> peruvian lily purple coneflower hard-leaved pocket orchid english marigold</p>	<p>petunia sweet william <u>pink primrose</u> tree mallow clematis</p>
	<p><u>giant white arum lily</u> moon orchid sweet pea canterbury bells hard-leaved pocket orchid</p>	<p><u>giant white arum lily</u> thorn apple moon orchid Lotus sweet pea</p>
	<p><u>english marigold</u> colt's foot yellow iris purple coneflower bird of paradise</p>	<p><u>english marigold</u> sunflower Gazania barbeton daisy yellow iris</p>
	<p>spear thistle bird of paradise globe thistle colt's foot <u>purple coneflower</u></p>	<p>artichoke bird of paradise blanket flower sunflower passion flower</p>
	<p>bird of paradise monkshood globe thistle <u>spear thistle</u> tiger lily</p>	<p>alpine sea holly toad lily artichoke sunflower orange dahlia</p>

Figure 7: AVSE classification results on **FLO** dataset. In the context of either ZSL or GZSL, the top-5 predictions are shown. Unseen classes are in green and seen classes are in blue. The prediction that is underlined represents the ground-truth class label. In the last image instance, the GZSL predictions fail to estimate the correct label.