Image-free Classifier Injection for Zero-Shot Classification

Anders Christensen[1,2] Massimiliano Mancini[3] A. Sophia Koepke[1]
Ole Winther[2,4,5,6] Zeynep Akata[1,7]
[1]University of Tübingen [2]Technical University of Denmark [3]University of Trento
[4]University of Copenhagen [5]Copenhagen University Hospital [6]FindZebra [7]MPI for Intelligent Systems
andchri@dtu.dk

Abstract

Zero-shot learning models achieve remarkable results on image classification for samples from classes that were not seen during training. However, such models must be trained from scratch with specialised methods: therefore, access to a training dataset is required when the need for zero-shot classification arises. In this paper, we aim to equip pre-trained models with zero-shot classification capabilities without the use of image data. We achieve this with our proposed Image-free Classifier Injection with Semantics (ICIS) that injects classifiers for new, unseen classes into pre-trained classification models in a post-hoc fashion without relying on image data. Instead, the existing classifier weights and simple class-wise descriptors, such as class names or attributes, are used. ICIS has two encoder-decoder networks that learn to reconstruct classifier weights from descriptors (and vice versa), exploiting (cross-)reconstruction and cosine losses to regularise the decoding process. Notably, ICIS can be cheaply trained and applied directly on top of pre-trained classification models. Experiments on benchmark ZSL datasets show that ICIS produces unseen classifier weights that achieve strong (generalised) zero-shot classification performance. Code is available at https://github.com/ExplainableML/ImageFreeZSL.

1. Introduction

With the immense growth of user-generated data, object categories are routinely discovered or re-defined, and requirements for models change constantly. Hence, pre-trained visual classifiers can rapidly turn inadequate as their output space is limited to classes seen during training. Updating a decision maker commonly requires expensive data collection and annotation, which can be unrealistic in some domains.

A cheaper and more appealing strategy is to use a zero-shot learning (ZSL) model [1, 57, 55, 56] that exploits side information (*e.g.* class label word embeddings) to recognise unseen classes in the specific domain of interest. However,

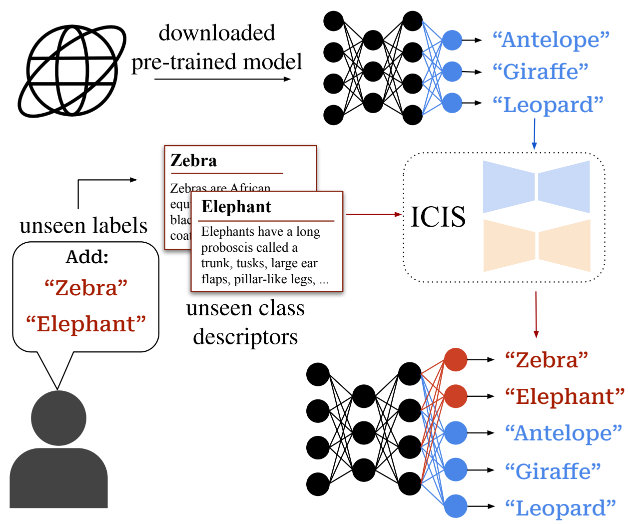


Figure 1: We propose Image-free Classifier Injection with Semantics (ICIS), a method that adds new classes to a pre-trained model *without* access to *any* image for seen and unseen classes, learning *only* from class-specific descriptors (*e.g.* attributes) and their corresponding classifier weights.

i) standard classification models are not designed with zero-shot capabilities [19], and ii) state-of-the-art ZSL methods are not applicable post training as one cannot regress descriptors from images [28, 1], train a feature generator [56], or learn visual-semantic embeddings [60, 54] without images from seen classes. Crucially, it can be infeasible for a practitioner to (re-)train a ZSL model due to computational costs or data storage issues, and some visual data might not be available for a particular task or user.

Since a large variety of pre-trained deep learning models are readily available online, we pose the question: given a specific image classification task and a pre-trained model, can we extend it to desired but missing categories *without* using images from seen or unseen classes? We name this task Image-free ZSL (I-ZSL), where the aim is to categorise samples from unseen classes without using image data, by injecting new classification weights into pre-trained classification models in a post-hoc manner. We note that while vision-language models (*e.g*. CLIP [43]) can perform zero-shot transfer, I-ZSL goes into an orthogonal direction by targeting models that are specialised for specific tasks, such as fine-grained recognition, for which model updates are needed when introducing new categories. In principle, an I-ZSL method could be combined with CLIP, *e.g*. by estimating class-specific prompts [40] for unseen classes.

We address the I-ZSL task by proposing Image-free Classifier Injection with Semantics (ICIS), a method that directly relates semantic class descriptors (*e.g*. class names or attributes) to the classifier weights of a pre-trained model. Concretely, ICIS receives as input classifier weights of seen classes and class descriptors. It learns two encoder-decoder networks, one for descriptor vectors and one for the classifier weights. These networks inject priors from the respective data sources into the simple descriptor-to-weights mapping. Moreover, we supervise the encoder representations by mapping across spaces (*i.e*. descriptor-to-weights and vice versa), using pairs of seen class descriptors and their corresponding classifier weights, in order to encourage better generalisation on unseen descriptors. Notably, ICIS is trained and used for classifier injection in a post-hoc manner, *i.e*. after the initial classification model has been trained and without access to the initial training data. We validate our approach on standard ZSL benchmark (*i.e*. CUB [49], SUN [41], AWA2 [55]), against applicable ZSL methods and adapted state-of-the-art zero- and few-shot learning approaches, consistently outperforming them in both generalised and standard evaluation.

To summarise, we make the following contributions: (1) We tackle the new image-free ZSL task, where zero-shot classification is performed by injecting new classification weights into a pre-trained classification model, without access to any samples this model was trained on. (2) We propose ICIS, a simple framework with two encoder-decoder architectures that directly predicts new classifier weights for unseen classes from limited training data, *i.e*. classifier weights for the seen classes and corresponding class-wise descriptors. (3) Despite the additional restrictions of the I-ZSL setting, ICIS achieves strong (Generalised)-I-ZSL performance on a variety of ZSL benchmarks, outperforming several zero- and few-shot approaches adapted to I-ZSL.

2. Related Work

Zero-Shot Learning. aims to classify samples from classes that were not part of the original training set [28], *e.g*. by using class-specific semantic information to relate seen and unseen classes. Commonly, methods learn to map images to class-level semantic descriptors [28, 1, 53, 45], improve visual-semantic embeddings [54, 32, 66, 6, 37], *e.g*. using part-aware training strategies [60, 59], or train generative models to synthesise visual features for unseen classes [56, 58, 46, 67, 68, 27, 18, 23, 7]. Recent methods have also considered novel ways of comparing visual features and semantic descriptions using visual patches [61] or noisy documents [38], and descriptions generated with GPT-3 [14, 3]. The aforementioned methods assume full access to the training data and the possibility to train a ZSL model from scratch. In this work, we remove this assumption and create a zero-shot classification model from a pre-trained classification one, *without* access to the images it was trained on, using only semantic descriptions of the classes.

Generating classifiers for ZSL. ICIS addresses the lack of training images in a direct way, learning a function that maps class descriptors to their corresponding classifier weights. In this regard, ICIS has some similarities to hypernetworks [17] and previous works that infer weights for zero-shot classification. Most of these rely on graph-convolutional neural networks (GCNs) [25]: [50, 21] predicts classifier weights given semantic descriptions from a class-hierarchy as side information. [50, 21, 16] use seen class images to refine the graph, fine-tune input features, and train a denoising autoencoder, respectively. Other approaches learn virtual classifiers for unseen classes [4] or sample images to train the classifier generator [30] when training the base classification model. Different from these approaches, ICIS does not use the original training data or scarcely available auxiliary information (*e.g*. class hierarchies). To the best of our knowledge, among previously published ZSL literature, only ConSE [39] and COSTA [36] can directly perform zero-shot classification without relying on images. While ConSE [39] maps images to the attribute space by a weighted combination of the seen class embeddings, COSTA [36] directly estimates classifier weights for unseen classes by exploiting their co-occurrence/similarity with seen ones. Despite the appealing simplicity of ConSE and COSTA, they do not explicitly consider the image-free setting, although being applicable for it.

Incremental, few-shot learning. Since ICIS extends a classification network to new classes, it is related to incremental [31, 26, 44, 13] and continual learning [64, 33, 42, 10], and in particular few-shot incremental learning (IL) [65, 12, 48, 2, 8, 20]. Interestingly, previous works also merged ZSL and IL [62, 52, 51, 22] to either improve ZSL models with a stream of data containing also unseen class images [62, 52], learn new attributes [22], or increase IL performance [51]. However, different from these works, we add new classes to the model without access to any image data of new or old classes. This makes our model insusceptible to catastrophic forgetting [15, 26], a common issue in IL.

Considering few-shot IL, our model is most closely related to [2], which explores different ways to predict classifier weights for new classes by using few-shot samples, side information, and subspace projections. Differently from [2], we have no images informing our descriptor-to-weights mapping and we do not rely on subspace regularisers. Instead, our ICIS learns to encode class descriptors and classifier weights into a shared embedding space whilst preserving their respective structures via reconstruction objectives.

3. Methodology

We first formalise the image-free ZSL task in Section 3.1. We then describe our method, ICIS, a simple, yet effective approach for I-ZSL (Section 3.2). Figure 2 provides an overview of our task and approach.

3.1. The Image-free ZSL task

In this work, we tackle the task of adapting a pre-trained classification model to categorise images of previously unseen classes, without access to any images.

Let us denote a given classification model as $\Phi : \mathcal{X} \rightarrow S$, where $\mathcal{X}$ is the image space, and S is the set of seen categories. Without loss of generality, we assume that Φ is composed of two modules, *i.e*. a feature extractor F and a classifier with weights W_S. We assume that Φ is pre-trained on an unavailable dataset with annotated seen class images. Our objective is to inject new classifiers into Φ to be able to classify images from a given set of target classes Y, *i.e*. producing the extended model $\hat{\Phi} : \mathcal{X} \rightarrow Y$. Note that we can have $Y \cap S = \emptyset$, as in the ordinary ZSL setup, or $Y \cap S = S$ if we wish to continue recognising seen classes, as in the more challenging case of GZSL. In the following, we will denote the set of unseen classes in Y as U, *i.e*. $U = Y \setminus S$.

To inject classifiers that extends Φ to U, we assume to have a set of vectors A, which describe the classes in S and U. We denote the semantic descriptors for classes in S and U as $A_S = \{\mathbf{a}_s\}_{s \in S}$ and $A_U = \{\mathbf{a}_u\}_{u \in U}$, with $A = A_S \cup A_U$. The only constraint for A is that all descriptors come from the same source (*e.g*. word embeddings, expert annotations), allowing us to model the relationships between seen and unseen classes. In principle, the set of descriptors can be obtained from generic linguistic sources (*e.g*. language models, word embeddings) and thus we can also use just embeddings of the class names as inputs.

From the above definition, the main challenge of image-free ZSL is the lack of *any* training images. This constraint prevents the use of standard ZSL techniques, such as generating features for unseen classes [56, 46] or learning specific compatibility functions [54, 60], since we cannot learn how visual inputs relate to the descriptor space. We instead work in a different space: the classification weights of the given pre-trained model. Specifically, we assume that the existing classifier weights W_S for the seen classes S already encode visual-semantic knowledge, and act as visual descriptors of the seen classes. From this assumption, we approach the task of image-free ZSL with the intention of mapping class descriptors to their corresponding classifier weights. We can then perform zero-shot classification by equipping Φ with inferred weights for unseen classes. In the following, we describe how we implement this mapping in ICIS.

3.2. Image-free Classifier Injection with Semantics

The Image-free Classifier Injection with Semantics (ICIS) framework for the I-ZSL extends a base model by predicting classifier weights from class-specific descriptors with an autoencoder structure and training objectives that regularise learning in this extremely data-scarce setting. Here, we first describe our base model and its extensions for I-ZSL.

Base model. Our goal is to produce a module mapping descriptors of unseen classes to corresponding classifier weights. Formally, let us denote the set of classifier weights of Φ for the seen classes in S by $W_S = \{\mathbf{w}_s\}_{s \in S}$. We want to learn a mapping $\phi_a^w : \mathcal{A} \rightarrow \mathcal{W}$ where $\mathcal{A}$ is the descriptor space and $\mathcal{W}$ the classifier weights space. The simplest approach to learn ϕ_a^w is by exploiting the available descriptors and classifier weights of the seen classes. With the latent space $\mathcal{Z}$, the descriptor encoder $E_a : \mathcal{A} \rightarrow \mathcal{Z}$, and the classifier weights decoder $D_w : \mathcal{Z} \rightarrow \mathcal{W}$, we can define $\phi_a^w = D_w \circ E_a$. We learn ϕ_a^w by minimising the following regression objective:

$$\mathcal{L}_{A \rightarrow W}^{S} = \sum_{s \in S} d(\phi(\mathbf{a}_s), \mathbf{w}_s), \tag{1}$$

where d is a generic distance function (*e.g*. mean-square error). During inference, we infer a classifier for an unseen target class $u \in U$ as:

$$\mathbf{w}_u = \phi_a^w(\mathbf{a}_u) = D_w(E_a(\mathbf{a}_u)), \tag{2}$$

and inject $\mathbf{w}_u$ into the resulting classification head W. If the downstream classification task is in the generalised zero-shot setting, we have $W = W_S \cup W_U$ with $W_U = \{\mathbf{w}_u\}_{u \in U}$, while for the ordinary zero-shot setting $W = W_U$.

The base model achieves good results in the zero-shot setting, but its performance rapidly decreases in the generalised case. In the following, we describe how ϕ_a^w can be improved to better generalise to unseen class descriptors in A_U.

Comparing vectors through cosine-similarity. The choice of the distance function d heavily affects the performance of the framework. While multiple choices are possible, we employ a simple implementation based on cosine-similarity:

$$d(\mathbf{v}, \mathbf{q}) = 1 - \cos(\mathbf{v}, \mathbf{q}) = 1 - \frac{\mathbf{v} \cdot \mathbf{q}}{||\mathbf{v}|| \cdot ||\mathbf{q}||}, \tag{3}$$

where $\mathbf{v}, \mathbf{q} \in \mathbb{R}^m$. Using the cosine distance, e.g. instead of L2 distance, allows the network to focus on the angular alignment of the injected weights with respect to the pre-trained ones, ignoring differences in magnitude. We found

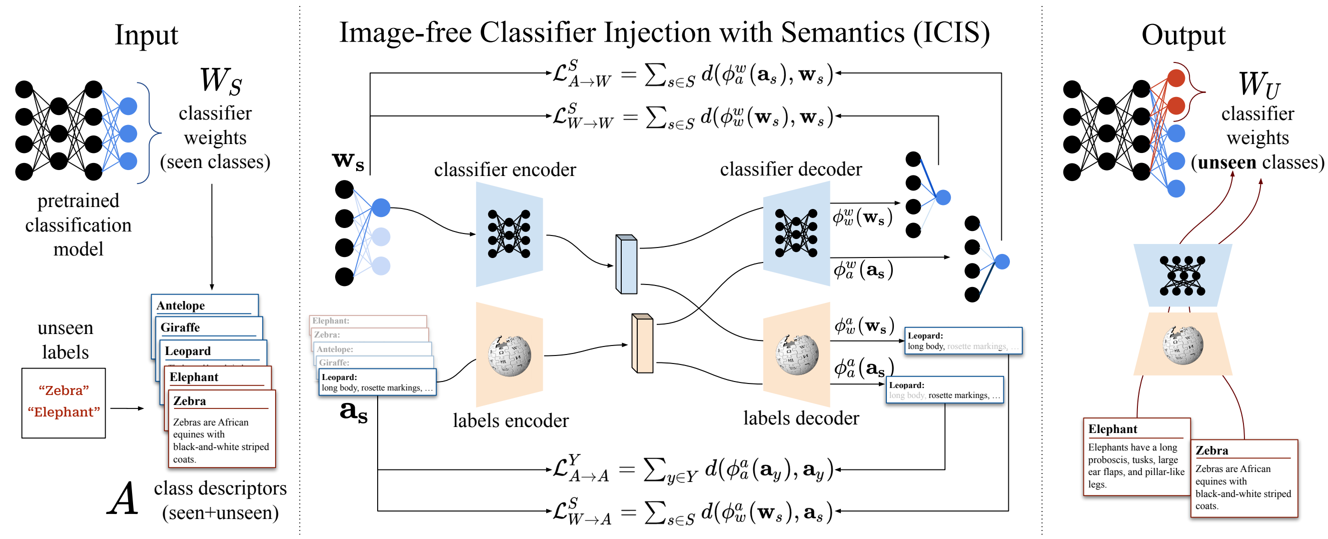


Figure 2: **Image-free Classifier Injection with Semantics (ICIS)**. Given a pre-trained model with classifier weights W_S trained on the set of seen classes S, our goal is to inject new classifier weights with weights for unseen classes (*e.g.* red nodes) into the set of existing weights W_S *without any images*, using only class-specific descriptors. To tackle this task, we instantiate two encoder-decoder architectures (blue and ocher) that map inputs within and across the two spaces, supervising them using pairs of descriptors and classifier weights for seen classes. During inference, we use the descriptor encoder and the weights decoder to infer weights for an unseen class given its descriptor. The inferred weights can then be injected into the set W_S (right), allowing the pre-trained model to recognise images of unseen classes.

the angular alignment to promote better generalisation and to reduce the usual bias on seen classes, common in GZSL [5].

Autoencoding to preserve semantic structure. Both the classifier weights in W_S and the descriptors in A are robust sources of visual/semantic information. Indeed, W_S has been trained on a dataset $\mathcal{T}$ with images of the seen classes, and consequently encodes distinctive visual relationships/similarities depicted in $\mathcal{T}$. At the same time, A might be derived from human expertise (*e.g.* animal descriptions) or from natural language resources trained on large corpora (*e.g.* word embeddings, language models). Given that we face an extreme low-data scenario, it would be ideal to regularise the descriptor-to-weights mapping ϕ^w_a by preserving the structure in the spaces $\mathcal{A}$ and $\mathcal{W}$.

We found that simply adding specific autoencoder structures for each of the spaces achieves this goal. Formally, we instantiate two additional networks $D_a : \mathcal{Z} \to \mathcal{A}$ mapping vectors in $\mathcal{Z}$ back to the descriptor space $\mathcal{A}$, and $E_w : \mathcal{W} \to \mathcal{Z}$ mapping classifier weights to the latent space $\mathcal{Z}$. We can then enforce the original structure of the descriptor space in ϕ^w_a by regularising E_a and minimising the following objective:

$$\begin{aligned} \mathcal{L}^S_{A\to A} &= \sum_{s\in S} d(\phi^a_a(\mathbf{a}_s), \mathbf{a}_s) \\ &= \sum_{s\in S} d(D_a(E_a(\mathbf{a}_s)), \mathbf{a}_s), \end{aligned} \tag{4}$$

where $\phi^a_a = D_a \circ E_a$. Note that in case we know the target unseen classes at training time, we can minimise the objective over the full set Y in Eq. (4), rather than just S. Similarly, we enforce that $\mathcal{Z}$ additionally preserves the structure of the given classifier weights of seen classes by minimising:

$$\begin{aligned} \mathcal{L}^S_{W\to W} &= \sum_{s\in S} d(\phi^w_w(\mathbf{w}_s)), \mathbf{w}_s) \\ &= \sum_{s\in S} d(D_w(E_w(\mathbf{w}_s)), \mathbf{w}_s), \end{aligned} \tag{5}$$

where $\phi^w_w = D_w \circ E_w$. With Eq. (4) and Eq. (5), we are enforcing that vectors encoded in $\mathcal{Z}$ can be reconstructed by their space-specific decoders. This encourages E_a and D_w to not only focus on the descriptor-to-weights mapping, but also to preserve the information encoded in their respective input/output spaces $\mathcal{A}$ and $\mathcal{W}$.

Latent space alignment. While the autoencoders preserve structure from the respective spaces, they do not ensure that the two latent spaces are aligned, *i.e.* a descriptor $\mathbf{a}_y$ and its corresponding classifier weights $\mathbf{w}_y$ may be distant in $\mathcal{Z}$. While vectors in the two spaces are already partially aligned by means of Eq. (1), we can exploit the two additional modules E_w and D_a to further impose alignment via a symmetric objective of mapping classifier weights to descriptors. This is achieved with the following term:

$$\mathcal{L}^S_{W\to A} = \sum_{s\in S} d(\phi^w_a(\mathbf{w}_s), \mathbf{a}_s) = \sum_{s\in S} d(D_a(E_w(\mathbf{w}_s)), \mathbf{a}_s), \quad (6)$$

where $\phi^a_w = D_a \circ E_w$. Combining this objective with the previous ones promotes the regularisation of ϕ^w_a toward aligned and structure-aware representations in $\mathcal{Z}$.

Full objective. Merging the four terms defined in the previous sections, our final objective becomes:

$$\mathcal{L} = \mathcal{L}^S_{A\to W} + \mathcal{L}^Y_{A\to A} + \mathcal{L}^S_{W\to W} + \mathcal{L}^S_{W\to A}, \quad (7)$$

where the first is our target mapping (Eq. (1)), the second and the third terms are the structure preserving ones within each of the spaces (Eq. (4) and Eq. (5)), and the last term is the reverse weights-to-descriptor objective, that encourages further alignment in $\mathcal{Z}$ (Eq. (6)).

4. Experiments

In this section, we provide experimental results of ICIS in the image-free ZSL task. We first describe the datasets, baselines used in our experiments, and training details of ICIS in Sec. 4.1. Experimental results of (generalised) zero-shot classification performed with injected weights from ICIS are showcased in Sec. 4.2), and Sec. 4.3 analyses the individual elements of our framework.

4.1. Experimental setting

Here, we describe the three datasets used for evaluating our framework ICIS and the baselines that we compare to.

Datasets. We evaluate I-ZSL performance on the standard ZSL benchmark datasets CUB [49], AWA2 [55], and SUN [41]. We use splits of seen and unseen classes proposed in [57, 55] for (generalised) zero-shot learning. CUB [49] is a dataset for fine-grained bird classification with 200 categories. Following [1, 55], we split the classes into 150 seen classes and 50 unseen classes: this means that our I-ZSL training set consists of 150 pairs of descriptors and classifiers. AWA2 [55] is a coarse-grained classification dataset 50 different animals (40 seen and 10 unseen classes). Finally, SUN is a dataset of scene classification consisting of 717 indoor and outdoor scenes. Following [29], we consider 645 classes as seen and 72 as unseen, using the split defined for (generalised) zero-shot learning in [55].

Baselines. To evaluate ICIS, we compare with baselines and existing methods in the literature that are compatible with [39, 36] or adaptable to [2, 16, 61] I-ZSL. *ConSE* [39] and *COSTA* [36] perform a weighted sum of existing seen class elements (embeddings the first, classifiers the latter) using as input either the current test sample (ConSE) or co-occurrence statistics (COSTA). *VGSE* [61] proposes to extract visual attributes from the seen classes, predicting the unseen class attributes via either weighted average of embedding similarities (*WAvg.*) or similarity matrix optimisation (*SMO*). We test these two strategies for I-ZSL, as they do not require seen class images.

Finally, we adapt the approaches of [2, 16] from the FSL literature. Subspace regularisers (*Sub. Reg.* [2]) encourage novel classifiers to be within the subspace spanned by the base classifiers of the given model, under semantic guidance by descriptors. wDAE [16] trains a GNN-based denoising autoencoder to improve initial estimate classifiers. In both cases, we train a simple MLP (with the same structure of ICIS) to provide candidate unseen class classifiers. We then apply the regularising projections on the predicted weights for Sub. Reg., while we use them as input to DAE-GNN instead of image feature means. In the supplementary material, we show results when combining these approaches with our ICIS, since the methods are complementary.

Hyperparameter selection without images. Since we have no access to any image data, we cannot select hyperparameters based on downstream classification performance on neither unseen nor seen classes. Instead, to determine suitable hyperparameters for our methods and all the baselines, we split the *descriptor-weights-pairs* of the seen classes into a training and validation split, using the same validation splits of seen classes proposed in [55, 57] for standard (generalised) zero-shot learning, without using any image.

While providing reliable hyperparameters, we found that using the validation loss as criterion for early stopping leads to poor convergence, due to our extremely low-shot learning setup. We sidestep this problem by training on all samples, and applying a simple stopping criterion based on the slope of the training curve. We provide more details on the hyperparameters and the criterion in the supplementary.

Implementation details. As the given pre-trained classification model, we use a ResNet101 [19] trained to classify only the seen classes in each respective dataset, as is common in the ZSL literature [58]. Across all datasets, the encoders and decoders of the simple ICIS framework are single layer linear mappings, with the encoder mappings being followed by a ReLU activation function. For CUB and AWA2, the hidden dimensionality is 2048, while it is 4096 for SUN. The mappings are trained with the Adam optimiser [24] with a learning rate of 10^{-5} and $(\beta_1, \beta_2) = (0.9, 0.999)$. We employ a batch size of 16 on CUB and SUN, while we set the batch size to 20 for AWA2.

4.2. Results on Image-free ZSL

In Table 1, we show experimental results of our ICIS and the competitors in I-ZSL, using the standard dataset-specific

Image-free Zero-Shot Learning	Zero-Shot Accuracy			Generalised Zero-Shot Accuracy								
	CUB	AWA2	SUN	CUB			AWA2			SUN		
	Acc	Acc	Acc	u	s	H	u	s	H	u	s	H
ConSE [39]	41.9	44.0	44.4	0.5	88.0	0.9	3.0	96.1	5.7	0.1	47.9	0.1
COSTA [36]	31.9	40.9	19.9	0.0	87.6	0.0	0.0	96.1	0.0	0.0	50.1	0.0
Sub. Reg.* [2]	37.6	37.5	48.3	0.0	87.6	0.0	0.0	96.1	0.0	0.0	50.1	0.0
wDAE* [16]	38.2	37.0	49.9	0.0	87.3	0.0	0.1	96.0	0.3	0.0	49.3	0.0
WAvg* [61]	2.0	20.1	1.4	1.9	52.3	3.7	5.5	92.4	10.4	0.0	50.1	0.0
SMO* [61]	45.1	55.4	42.7	39.2	52.3	44.8	31.8	92.4	47.3	42.5	1.6	3.1
ICIS (Ours)	**60.6**	**64.6**	**51.8**	45.8	73.7	**56.5**	35.6	93.3	**51.6**	45.2	25.6	**32.7**

Table 1: Comparison between our proposed framework ICIS and existing methods in the literature applicable or adaptable to the image-free ZSL (I-ZSL) setting using standard ZSL benchmarks (*i.e*. CUB, AWA2, and SUN). We measure the results as unseen accuracy (Acc) for the zero-shot task, unseen (u) and seen (s) accuracy and their harmonic mean (H) for the generalised zero-shot setting. Methods marked with * are adapted to the image-free setting.

ICIS Ablation	Zero-Shot Accuracy			Generalised Zero-Shot Accuracy								
	CUB	AWA2	SUN	CUB			AWA2			SUN		
	Acc	Acc	Acc	u	s	H	u	s	H	u	s	H
MLP base model	41.4	46.8	49.7	0.0	87.6	0.0	2.0	95.9	4.0	0.0	50.1	0.0
+ Cosine loss	52.7	50.9	48.5	36.6	76.6	49.5	27.3	93.1	42.2	21.9	43.6	29.2
+ Within spaces	58.0	63.5	50.8	39.6	77.6	52.5	33.8	93.5	49.6	38.9	32.5	35.4
+ Across spaces	60.1	64.8	51.7	44.5	75.0	55.9	35.5	93.6	51.4	44.0	27.6	33.9
+ Include A_U	60.6	64.6	51.8	45.8	73.7	56.5	35.6	93.3	51.6	45.2	25.6	32.7

Table 2: Ablation study of the individual elements of our proposed ICIS. Starting from the MLP base model with standard L2 loss, we replace the latter with the cosine distance. We then analyse the encoder-decoder architectures by performing mappings within and across spaces. Finally, we check the impact of adding unseen class descriptors during training.

attributes adopted in ZSL [55] with dimensionalities 312, 85, and 102 for CUB, AWA2, and SUN, respectively. This makes the results more comparable to those from ordinary ZSL approaches that have access to images.

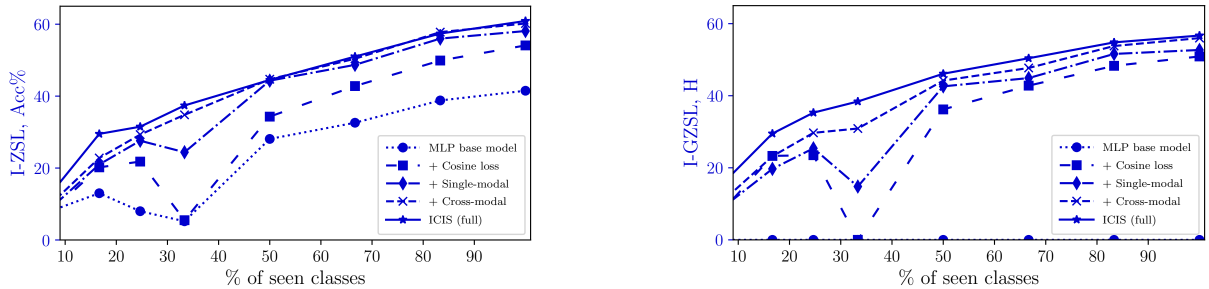
As the table shows, ICIS consistently outperforms all competitors, both in standard and generalised ZS classification. When only unseen classes are predicted (zero-shot accuracy), ICIS improves over the best ZS baselines by a large margin in all settings (*i.e*. +15.5% on CUB and +9.2% on AWA2 over SMO, and +7.4% over ConSE on SUN). The same applies to the adapted FS baselines, with ICIS outperforming the best method by more than 20% on both CUB and AWA2, and by 2% on SUN.

These margins are even more evident in the generalised setting, where ICIS is the only method that consistently provides a good trade-off between accuracy on seen and unseen classes (*e.g*. achieving a harmonic mean of 56.5 on CUB), while all other baselines but SMO fail in this setting. This is due to their inherent bias toward seen categories that our method largely alleviates. These results demonstrate that ICIS can successfully inject inferred classifiers for unseen classes into an existing classification network without relying on image training data.

4.3. Analyses of our ICIS framework

In this section, we analyse the contributions of the individual components of our ICIS framework, ablating the impact of each technical component in this extremely low-shot setting, and how the choice of embeddings and pre-training affects ICIS performance.

Ablation study. Here, we validate the benefit of the main components of our proposed model, showing the results in Table 2. We start with the MLP base model, that maps descriptors to classifier weights and it is trained with a standard L2 loss. This model achieves good results on ZS but its performance drastically decreases in GZSL, since the resulting classifiers tend to overly predict seen classes. Changing the objective to a cosine loss (+ *Cosine loss*) largely mitigates this bias problem (*e.g*. 49.5 harmonic mean on CUB, with 36.6% unseen accuracy vs 0.0% of the base model). The angle-based loss leaves the norm of the output weights to be implicitly regularised, but improves compatibility of the new predicted classification weights with the existing ones for seen classes, largely increasing the entropy on the prediction between the two sets (*e.g*. from 0.89 to 2.8 on CUB, see the additional analysis in the supplementary).



(a) Impact on I-ZSL performance, unseen Acc%.

(b) Impact on I-GZSL performance, harmonic mean H.

Figure 3: Data efficiency of architecture ablations of ICIS on the CUB dataset. Over the various dataset sizes, each addition to the MLP base model up to the full ICIS results in increasing performance for both the zero-shot and generalised zero-shot task.

With '*Within spaces*' in Table 2, we refer to the intermediate step of including the reconstruction losses on class descriptors (Eq. (4)) and classifiers (Eq. (5)). This leads to improved results on all metrics and datasets, *e.g*. from 50.9% to 63.5% unseen accuracy on AWA2 for ZS, and from 29.2 to 35.4 harmonic mean on SUN GZSL. We ascribe this to the implicit regularisation added by forcing the embeddings to retain semantic knowledge, such that the class-specific distinctive attributes can again be decoded from them.

Including the reconstruction objective from weights to class descriptors Eq. (6) leads to improved results everywhere, but on the SUN seen class accuracy. This additional alignment objective, matching latent embeddings across spaces, appears to be otherwise beneficial for the extreme low-data setting of the considered datasets.

Finally, including descriptors of unseen classes (+ Include A_U) in the descriptor-to-descriptor mapping during training (Eq. (4)) leads to marginal improvements, suggesting that while unseen class information may help, we can still train ICIS without a-priori knowledge of the target unseen classes.

Data-scarcity and generalisation. Here, we further examine how the components of ICIS combat overfitting and improve generalisation in this data-scarce setting. In Fig. 3, we show results for training our framework only with subsets of the descriptors of seen classes in the CUB dataset. We note that when training ICIS on a lower number of descriptor-classifier pairs, we still evaluate the generalised zero-shot classification accuracy by injecting the predicted classifiers into the classification model with the full set of seen classes.

We observe that when training ICIS on dataset sizes varying from just 12 descriptor-classifier pairs to the full training set of 150 pairs, the performance of the classifiers predicted by ICIS on both zero-shot classification and the generalised task increases consistently with each added component. In general, Fig. 3 confirms the trends of Table 2, showing that each component improves the robustness to the number of training descriptor-classifier weights pairs in the extremely constrained I-ZSL. Notably, when ICIS has as little as 50% of seen class descriptors, it already outperforms the MLP base model with full availability (Fig. 4).

Analysis of a failure case: *Tree Sparrow*. In this section, we study some shortcomings of ICIS on a concrete example considering the CUB dataset. In the generalised I-ZSL, on the unseen class *Tree Sparrow* our model regresses classification weights which correctly classify only 5% of the corresponding samples: this is the lowest accuracy our model achieves on an unseen class in this setting. We investigate this issue by counting which classes the *Tree Sparrow* samples were misclassified with, ordering them by similarity with the target classes (as computed from their class descriptors). Since the majority of the predictions are located within the closest 10 classes (see supplementary), in Fig. 4 we show only these classes, ordered by similarity.

As the figure shows, the injected classifiers from ICIS result in classifying primarily the most similar classes to *Tree Sparrow* - and mostly within the sparrow family. This suggests that the injected weights indeed classify based on properties close to those of *Tree Sparrow*, showing that ICIS captured its core visual properties. Among the top predicted classes, we can see that there are both seen and unseen ones, showing that our model is not overly biased toward seen classes. Indeed, misclassifications among seen classes only occurs on the two most similar classes to *Tree Sparrow*, namely *Chipping Sparrow* and *House Sparrow*, that apart from extremely fine-grained details (*e.g*. beak shape and size in *House Sparrow*, see examples in the supplementary) are difficult to distinguish from the target class, in which cases the injected classifiers may fall short. Nevertheless, this analysis confirms that ICIS has learned from the visual semantic knowledge encoded in existing classifier weights: this is reflected on the injected weights for new, unseen, classes, being able to capture discriminative visual properties.

4.4. I-ZSL from ImageNet

In this section, we analyse how the ZS performance of ICIS and the competitors change when i) I-ZSL is performed

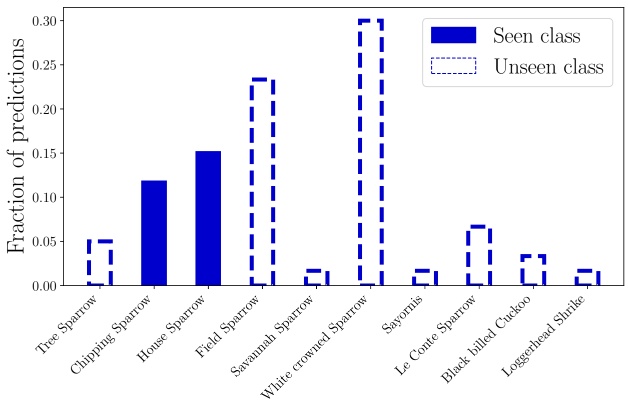


Figure 4: Predicted classes in the failure case *Tree Sparrow* (CUB, I-GZSL), where the x-axis indicates predicted CUB classes ranked by attribute similarity to *Tree Sparrow*, while the y-axis indicates the times an image of *Tree Sparrow* is assigned to a class. The misclassifications are distributed mostly between classes of other unseen (dashed) Sparrow-classes, as well as the two seen classes (solid) that are closest to *Tree Sparrow* in terms of attribute similarity.

from a generic pre-trained object classification model and ii) how the performance changes w.r.t. the available class descriptors. For the former, we consider a ResNet101 [19] pre-trained on ImageNet [11], since it is readily available online via libraries such as `torchvision` [35]. For the embeddings, we consider Wikipedia word embeddings from [63], ConceptNet Numberbatch [47] and CLIP text embeddings [43] that also encode visual knowledge. These embeddings have dimensionality 300, 300, and 512, respectively.

The results are shown in Table 3. ICIS consistently achieves the top results across dataset and class descriptors, with the exception of Wiki2Vec and ConceptNet in CUB. These results, coupled with the ones in Table 1, highlight the robustness of our method to the particular pre-trained classification model and the available embeddings, making ICIS the most effective solution for image-free ZSL.

From the table, there are also some general observations that can be drawn. Results are much lower than the I-ZSL ones in Table 1 for SUN and CUB in particular (*e.g.* -13.7% and -22.5% for ICIS with CLIP embeddings). Especially on CUB, the performances are very low for non-visually informed embeddings (*e.g.* 6.0% for wiki2vec, 10.2% for ConceptNet). Since CUB is the dataset with the highest distribution shift w.r.t. ImageNet classes, these results highlight the challenges of I-ZSL when seen and unseen classes largely differ. In such settings, non-trained approaches (*e.g.* ConSE) can even outperform trained models (*e.g.* ICIS, wDAE) when class descriptors are less discriminative. On the other hand, on AWA2 the downstream performance of ICIS when training it on a pre-trained ImageNet model can even outperform

Model	**CUB**			**AWA2**			**SUN**		
	WV	**CN**	**CL**	**WV**	**CN**	**CL**	**WV**	**CN**	**CL**
ConSE [39]	3.8	**11.6**	11.7	50.3	60.9	72.9	17.2	25.4	17.8
COSTA [36]	3.9	1.6	14.1	53.4	51.8	54.9	12.5	2.6	18.8
Sub.Reg.* [2]	3.7	4.6	4.5	18.7	60.0	10.1	3.7	25.0	1.6
wDAE* [16]	4.0	6.0	26.9	64.4	59.6	84.5	17.8	24.6	31.8
WAvg* [61]	5.5	4.6	12.2	56.7	58.8	62.5	11.9	14.4	6.5
SMO* [61]	**7.3**	11.1	12.7	52.4	64.5	54.7	16.4	20.2	13.1
ICIS (Ours)	6.0	10.2	**27.8**	**65.2**	**64.6**	**86.1**	**19.7**	**25.9**	**38.1**

Table 3: Comparison between our ICIS and existing methods in the image-free ZSL (I-ZSL) setting from ImageNet to the set of unseen classes of ZSL benchmarks. We also test various types of class descriptors, *i.e.* Wiki2Vec [63] (WV), ConceptNet [47] (CN), and CLIP [43] text embeddings (CL). We measure the results as unseen accuracy for the zero-shot task. Methods marked with * are adapted to I-ZSL.

ICIS on a model trained on the seen classes from AWA2 specifically, reaching zero-shot accuracies up to 86.1%. This is not surprising as the seen classes of AWA2 have a large overlap with the ImageNet classes, and that the number of descriptor-classifier pairs (and hence the training dataset for ICIS) is 25 times larger (*i.e.* from 40 to 1000). We note that the unseen target classes are disjoint from the ImageNet classes when using the split proposed in [55]. The results on AWA2 and SUN demonstrate that it is possible to approach I-ZSL even when only a generic classifier is available, opening the possibility of constructing classification networks suited for tasks for which no visual dataset is available.

5. Conclusion

In this work, we presented the task of image-free ZSL (I-ZSL) of pre-trained models which generalises zero-shot learning to a setting in which access to any per-sample information (such as images or image features) is not possible. We present a simple framework, Image-free Classifier Injection with Semantics (ICIS) that allows for adapting any given classifier by predicting new classifier weights from semantic per-class descriptors (*i.e.* attributes, word embeddings). ICIS improves generalisation by regularising the mapping from descriptors to weights with additional mappings within and across the corresponding spaces, injecting semantic priors in the model. Our framework is computationally efficient and works under minimal assumptions and supervision, *i.e.* only classifier weights of the seen classes and their respective class descriptors. Our simple approach surpasses by a large margin all competitors, especially in the challenging generalised zero-shot classification task. ICIS consistently outperforms competitors adapted from ZSL and FSL literature, demonstrating that the extremely constrained image-free ZSL for pre-trained models can be tackled effectively, opening avenues for future research in this topic.

Acknowledgements. This work was supported by DFG project number 276693517, BMBF FKZ: 01IS18039A, ERC (853489 - DEXIM), EXC number 2064/1 – project number 390727645, and by the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU. Ole Winther is supported by the Novo Nordisk Foundation (NNF20OC0062606) and the Pioneer Centre for AI, DNRF grant number P1. Anders Christensen thanks the ELLIS PhD program for support.

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE PAMI*, 2016.

[2] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. In *ICLR*, 2021.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016.

[5] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016.

[6] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *CVPR*, 2022.

[7] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, 2021.

[8] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *CVPR*, 2021.

[9] Yu-Ying Chou, Hsuan-Tien Lin, and Tyng-Luh Liu. Adaptive and generative zero-shot learning. In *ICLR*, 2020.

[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE PAMI*, 2021.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[12] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *AAAI*, 2021.

[13] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, 2020.

[14] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *CVPR*, 2023.

[15] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[16] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. *CVPR*, 2019.

[17] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[18] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021.

[19] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[20] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *CVPR*, 2022.

[21] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, 2019.

[22] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, 2012.

[23] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *CVPR*, 2020.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[26] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017.

[27] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *CVPR*, 2022.

[28] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[29] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE PAMI*, 2013.

[30] K. Li, Martin Renqiang Min, and Yun Raymond Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *ICCV*, 2019.

[31] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE PAMI*, 2017.

[32] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration

network. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[33] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *NeurIPS*, 2017.

[34] Maren Mahsereci, Lukas Balles, Christoph Lassner, and Philipp Hennig. Early stopping without a validation set. *ArXiv*, abs/1703.09580, 2017.

[35] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *ACM MM*, 2010.

[36] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[37] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *CVPR*, 2020.

[38] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. In *NeurIPS*, 2022.

[39] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Gregory S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[40] Sarah Parisot, Yongxin Yang, and Steven G. McDonagh. Learning to name classes for vision and language models. *ArXiv*, abs/2304.01830, 2023.

[41] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 2014.

[42] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, 2020.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[44] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

[45] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[46] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *cvpr*, 2019.

[47] Robyn Speer and Joanna Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *International Workshop on Semantic Evaluation (SemEval-2017)*, 2017.

[48] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *CVPR*, 2020.

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[50] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *CVPR*, 2018.

[51] Kun Wei, Cheng Deng, Xu Yang, and Maosen Li. Incremental embedding learning via zero-shot translation. In *AAAI*, 2021.

[52] Kun Wei, Cheng Deng, Xu Yang, and Dacheng Tao. Incremental zero-shot learning. *IEEE Transactions on Cybernetics*, 2021.

[53] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[54] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, 2019.

[55] Yongqin Xian, H. Christoph Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE PAMI*, 2018.

[56] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. *CVPR*, 2018.

[57] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017.

[58] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019.

[59] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *ECCV*, 2020.

[60] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020.

[61] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *CVPR*, 2022.

[62] Nan Xue, Yi Wang, Xin Fan, and Maomao Min. Incremental zero-shot learning based on attributes for image classification. In *ICIP*, 2017.

[63] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *EMNLP*, 2020.

[64] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017.

[65] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. *CVPR*, 2021.

[66] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. *CVPR*, 2017.

[67] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. *CVPR*, 2018.

[68] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and A. Elgammal. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. *ICCV*, 2019.