SemGIF: A Semantics Guided Incremental Few-shot Learning Framework with Generative Replay

S Divakar Bhat[1]
sdivakarbhat@gmail.com
Biplab Banerjee[2]
getbiplab@gmail.com
Subhasis Chaudhuri[1]
sc@ee.iitb.ac.in

[1] Department of Electrical Engineering
Indian Institute of Technology Bombay

[2] Centre of Studies in Resources Engineering
Indian Institute of Technology Bombay

Abstract

We address the problem of incremental few-shot learning (IFSL) by leveraging the notion of generative feature replay. Learning novel concepts while preserving old knowledge is a long-lasting challenge in machine learning. The main concern in IFSL is to combat the catastrophic forgetting of the base classes whose training data are not available during the incremental stage while ensuring good generalization for the few-shot classes. Existing techniques prefer to preserve some base class samples to tackle forgetting, which does not comply with the intention of incremental learning. To this end, we propose a novel framework called Semantics Guided IFSL (SemGIF), which trains a generative model to synthesize base class samples on demand during the incremental step. Considering the importance of modeling a discriminative feature space in IFSL for separating the base and the novel classes, we propose a feature augmentation strategy where the visual embeddings are supplemented with the semantic features obtained from a word-embedding space. Such a feature space is found to produce enriched class prototypes to be utilized during classification. Experimental results on CIFAR-100, CUB, mini-ImageNet, and tiered-ImageNet in the homogeneous (within-dataset) and a novel heterogeneous (cross-dataset) setup showcase sharp improvements than the literature.

1 Introduction

Deep learning is widely adopted into several application areas due to its superior generalization ability from large-scale training databases. Notwithstanding the above, the data collection and annotation process are often tedious and costly. As a possible remedy, the research community is devoted to training deep learning models from less training data while ensuring that the trained models do not overfit [6, 7]. On a different note, it is crucial to develop inference models to learn continually as and when labeled data are made accessible. The notion of incremental or continual learning [26, 37] is deduced to handle such non-stationary setups where the model is required to continuously adapt to new tasks while judiciously controlling the catastrophic forgetting for the past tasks.

Recently, considerable progress can be observed in the field of few-shot learning (FSL) [7] which aims to train a network to learn from a small number of examples. Under this premise, the traditional FSL setup consists of base classes with voluminous labeled samples during training and a disjoint set of novel classes with a few training samples during testing, and the performance is always evaluated on the few-shot classes. This strategy has two possible bottlenecks. First, the discriminative information of the base classes is compromised, which would otherwise offer good separability for the few-shot categories. Secondly, humans can quickly learn novel concepts from limited experience on top of the existing knowledge base. Similarly, it would be interesting to design deep learning models which would attempt to handle novel categories while retaining the ability to classify a set of original base classes.

In view of the above, we understand that few-shot learning models that perform equally well on the base and novel classes are much desired. In this line, [22, 25] introduced the problem of IFSL where a backbone model is pre-trained on the base classes under abundant supervision, following which novel classes with little training data appear. The base class samples may not be available at this point due to memory constraints or security issues. In this situation, the IFSL model needs to combat the forgetting of the base classes while ensuring good generalization for the novel classes. Achieving this objective is non-trivial, and the direct extension of the FSL models does not work. The few existing works for IFSL focus on generating the classification weights for the novel classes directly from the base model [27]; however, they fail to generate discriminative class prototypes for the novel classes as the feature backbone is biased to the old task. It is also worth mentioning that a related but different problem setup from IFSL is few-shot class incremental learning (FSCIL) [1, 19, 36, 46] consisting of a large number of incremental episodes with each stage accumulating knowledge from a small number of classes under limited supervision. Moreover, the FSCIL problem is also primarily different from IFSL due to the former being originally a class incremental learning problem with the constrain of availability of a small number of labeled samples in each step. While the latter is fundamentally an FSL problem extended to incrementally accomplish more than the traditional FSL in terms of the ability to generalize over the base and novel dataset.

It is recently found in the class incremental learning literature that the replay-based approaches can better handle forgetting than the regularization-based methods like in [4, 26]; however, its effects are yet to be formalized judiciously for the IFSL task. From another point of view, the low-shot learning literature [15, 16] has showcased the importance of a semantic space for modeling better embeddings. Learning a discriminative space from limited visual data has been a long-lasting issue in FSL in general, and we are interested in finding the possibility of using semantic side information for improved feature learning in IFSL.

Inspired by the discussions, we propose a novel IFSL framework called SemGIF, which uses a generative model as the replay memory for the base classes while class semantics are used to obtain an improved embedding space by fusing the visual and the semantic information. Since the semantic information is generally class discriminative, the new embedding space shows less variance than the prototypes obtained only with the visual features. SemGIF follows a two-stage training protocol: in the pre-training stage, a conditional GAN is trained on the base samples, a visual to semantic mapping module is trained, and both the features are combined to make a new embedding space. The incremental stage uses the pre-trained semantic mapping module to train the model on the combined set of synthetic base and novel class samples. We summarize our significant contributions as,

i) We propose SemGIF, a novel IFSL framework based on a generative replay and the learning of a semantically influenced embedding space. Specifically, we introduce a se-

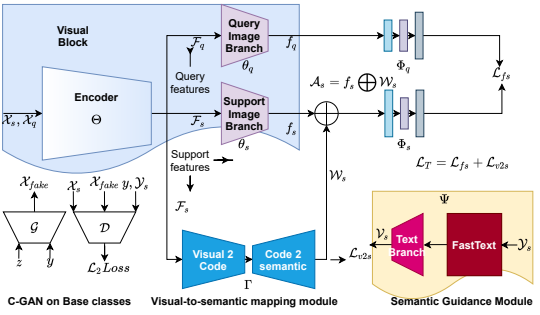


Figure 1: The complete pipeline of the proposed SemGIF framework depicting the visual block, the visual to semantic mapping module and the semantic guidance module. The embedding functions and the conditional GAN can also be seen. Details can be found in 3.2.

mantic embedding-based feature augmentation strategy for learning better class prototypes which is the first of its kind to be explored in the IFSL setting. We are also the first to explore generative replay in the IFSL domain to the best of our knowledge. ii) Unlike the previous works, we showcase the efficacy of SemGIF for two distinct experimental scenarios: homogeneous where the base and novel classes are obtained from the same distribution (dataset) and heterogeneous or dataset incremental FSL where the base and few-shot classes arise from different datasets. iii) We find SemGIF to be superior through rigorous ablations on all the four benchmark datasets. For instance, the performance boost is significant, with SemGIF outperforming the closest performing algorithm by about 8.63% and 24.51% in the case of, say, 5-way 5-shot on standard datasets like miniImagenet and CUB-200, respectively.

2 Related Works

Incremental Learning: The prime goal in incremental learning is to mitigate the effects of catastrophic forgetting [20, 21]. This forgetting can be controlled by resorting to memory-based approaches as detailed in the works [22, 26]. [17] proposed to use distillation loss to retain knowledge corresponding to past tasks. Regularisation can also be used to control the effect of forgetting without retaining the old samples as seen in the work [14]. Alternatively, rehearsal based approaches [4, 26, 41] relied on retaining a small number of exemplar samples or resorted to techniques that generate synthetic images [23, 33] or features [11, 43].
Few-shot learning: [6, 7] proposed a new machine learning paradigm called Few-Shot Learning (FSL) in order to facilitate learning from limited supervised information. The impact of alleviating the data gathering effort can be widely seen from the plethora of fields benefited from the same like image recognition [39], image retrieval [38] and gesture recognition [24]. Taking into account the relevance of the topic in many domains, there has been several machine learning approaches proposed in this direction, like embedding learning [2, 35, 39], meta-learning [8, 31], generative modeling [5, 30]. In this regard, one of the popular approaches is [34] which introduced a prototypical network to learn an embedding space and achieved promising results. There exist several extensions of this model; for example, [12] combined prototypical learning along with variational inference to learn a continuous embedding space.
Incremental Few-shot Learning: The principal objective of IFSL is to learn to classify novel samples while at the same time preserving the information acquired on the base classes, given that the novel samples only offer a small set of labeled samples. Most of the prior

works tackle this problem by generating the classification weights for novel classes while the pre-trained backbone network weights on base classes are fixed. In [25], novel class prototypes were directed to be used as the classification weights for both the base and novel categories. While [9] proposed to learn novel classification weights using a meta-learning weight generator which is fed with the novel class prototypes and base class weights. The attention attractor network of [27] utilized attention-based regularisation to prevent forgetting. Very recently, [45] proposed to extract representations for novel samples and perform efficient task-conditioning of base and novel classifiers by utilizing a task adaptive representation. SemGIF is entirely different from these methods as we explicitly utilize rehearsal-based feature replay and an additional semantic space while designing the class prototypes.

3 Methodology

3.1 Problem statement and preliminaries

In a standard N way K shot FSL problem the model is meta trained on the base dataset $\mathcal{D}_{base} = \{\mathcal{X}_{seen}, \mathcal{Y}_{seen}\}$ with N_b number of semantic categories, and further finetuned & evaluated on the few-shot dataset $\mathcal{D}_{novel} = \{\mathcal{X}_{unseen}, \mathcal{Y}_{unseen}\}$ with N classes and K labelled samples per class. The seen and unseen classes are disjoint in nature: $\mathcal{Y}_{seen} \bigcap \mathcal{Y}_{unseen} = \phi$. Inspired from [34], the support set $\{\mathcal{X}_s, \mathcal{Y}_s\}$ with $\mathcal{X}_s = \{X^1, \ldots, X^N\}$ where the cardinality of each X^i is K and the query set, $\mathcal{X}_q$, respectively work as the training and the validation sets used to learn a generic few-shot classification model by mimicking the testing scenario.

Even though we follow a similar episodic approach, the IFSL paradigm is alternatively tested on the entire $\mathcal{Y}_{seen} \bigcup \mathcal{Y}_{unseen}$ consisting of $N + N_b$ categories. Thus, for the evaluation of the model ($\hat{\Theta}$) on a joint prediction over both the base and novel dataset, a mini-batch of unlabelled query set is sampled in every episode, $\mathcal{Q} = \mathcal{Q}_{seen} \bigcup \mathcal{Q}_{unseen}$, such that the mapping, $\hat{\Theta}(\mathcal{Q}) \longrightarrow \mathcal{Y}_{\mathcal{Q}}$ would satisfy, $\mathcal{Y}_{\mathcal{Q}} \subset \mathcal{Y}_{seen} \bigcup \mathcal{Y}_{unseen}$.

3.2 Semantics Guided Incremental Few-shot Learning Framework

1. Model overview: The model ($\hat{\Theta}$) consists of a visual block, a semantic guidance module (Ψ), and a learnable visual to semantic mapping block (Γ) (Figure 1). The visual block consists of Resnet-18 backbone encoder (Θ), the support image branch (θ_s) and query image branch (θ_q) for the labeled support set ($\mathcal{X}_s$) and the unlabelled query sample set ($\mathcal{X}_q$), respectively. Our overall objective is to learn an efficient visual encoder ($\Theta + \{\theta_s, \theta_q\}$) and embedding functions ϕ_q, ϕ_s with the guidance from the support set $\mathcal{X}_s$. Note that ϕ_s and ϕ_q are responsible for mapping the features of each class of the support and query sets to an embedding space where the samples from each class tend to cluster around a representative class prototype. We use the shorthand, $\hat{\Theta} \longleftarrow \Theta + \{\theta_s + \{\phi_s\}, \theta_q + \{\phi_q\}\}$ to denote the resultant model obtained after training.

2. Training and inference: The overall training process shown in the Algorithm 1, can be split into two stages: a pre-training stage using $\mathcal{D}_{base}$ on the classification objective and an incremental stage once $\mathcal{D}_{novel}$ is introduced. Amongst different model components, we note that Γ is trained on the pre-training stage, which is directly used to approximate the semantic embeddings for the novel classes during the incremental stage. The complete flow of both the pre-training stage and the incremental stage is shown in Figure 2. We follow an episodic training scheme for both the stages where we generate an episode by randomly sampling

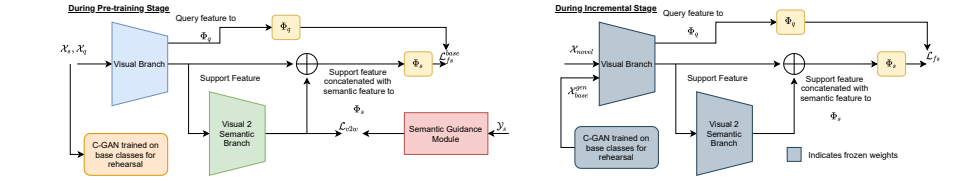


Figure 2: Illustration of flow of the proposed method. The figure on the left corresponds to the pretraining stage while that on the right depicts the active blocks in the incremental stage.

K labeled samples to act as the support set $\mathcal{X}_s$ over some randomly selected N classes from $\mathcal{D}_{base}$, while the remaining class-specific samples are considered to form the query validation set $\mathcal{X}_q$. In the incremental stage, we tune the model again in an episodic manner, over the novel dataset $\mathcal{D}_{novel}$ and the synthetic base samples generated using the C-GAN to mitigate forgetting since $\mathcal{D}_{base}$ is unavailable during the incremental stage. Both these stages, the visual-to-semantic mapping module, the augmented support features, and the C-GAN for the replay, are discussed below, along with the cost functions involved.

a. Pre-training stage: During the pre-training stage over the base dataset $\mathcal{D}_{base}$, we pass the set of query and support set samples through the encoder Θ to obtain the visual feature vectors $\mathcal{F}_q$ and $\mathcal{F}_s$, respectively. Subsequently, the support set features $\mathcal{F}_s$ are fed to the visual to semantic mapping module Γ to obtain $\mathcal{W}_s = \Gamma(\mathcal{F}_s)$. Γ is trained by minimizing the visual to semantic conversion loss $\mathcal{L}_{v2s}$ as shown in Equation 2. The output of the visual to semantic mapping module $\mathcal{W}_s$ is then concatenated with the corresponding visual support features, $f_s = \theta_s(\mathcal{F}_s)$ to form the new augmented support feature, $\mathcal{A}_s = f_s \bigoplus \mathcal{W}_s$, which then acts as the input to the embedding function $\phi_s : \mathbf{R}^M \longrightarrow \mathbf{R}^D$ ($\bigoplus$ denotes concatenation operation), where $M = |f_s| + |\mathcal{W}_s|$. Simultaneously, we also obtain the output $\phi_q(f_q)$ of the embedding function $\phi_q : \mathbf{R}^m \longrightarrow \mathbf{R}^D$ corresponding to the unlabelled query set, where $f_q = \theta_q(\mathcal{F}_q)$ and $m = |f_q|$. We seek to minimize the distance between the query samples and corresponding class prototypes μ obtained from the augmented support features as explained in Equation 3 thus facilitating ϕ_q to learn better embedding for the unlabelled query set. This is achieved by minimizing the few shot loss function $\mathcal{L}_{fs}$ given in Equation 4. The total cost function employed in this stage is shown in Equation 1.

Algorithm 1: The complete SemGIF framework.

Require: $\mathcal{D}_{base}, \mathcal{D}_{novel}, \Theta, \theta_s, \phi_s, \theta_q, \phi_q, \Gamma, \Psi$
Ensure: $\hat{\Theta} \longleftarrow \Theta + \{\theta_s + \{\phi_s\}, \theta_q + \{\phi_q\}\}$
while *iteration* $<$ *max iteration* **do**
  Get episode $\mathcal{B} \subset \mathcal{D}_{base} \longrightarrow \{\{\mathcal{X}_s, \mathcal{Y}_s\}, \{\mathcal{X}_q\}\}$
  $\mathcal{F}_q \longleftarrow \Theta(\mathcal{X}_q), \mathcal{F}_s \longleftarrow \Theta(\mathcal{X}_s)$
  $\mathcal{L}_{v2s} \longleftarrow \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K} \|\Psi(\mathcal{Y}_s) - \Gamma(\mathcal{F}_s)\|_2^2$ `Refer Eqn 2`
  $\mathcal{L}_{fs} \longleftarrow -\log p(y = k|x_q)$ `Refer Eqn 4`
  $\mathcal{L}_T \longleftarrow \mathcal{L}_{fs} + \mathcal{L}_{v2s}$
  Update: $\hat{\Theta} = \hat{\Theta} - \alpha\nabla\mathcal{L}_T$
  Train C-GAN as per Equation 6
end
Incremental learning stage
$\Phi \longleftarrow \{\phi_s, \phi_q\}$
while *iteration* $<$ *max iteration* **do**
  Get episode $\mathcal{B} \longrightarrow \{\{\mathcal{X}_s, \mathcal{Y}_s\}, \{\mathcal{Q}\}\}$
  $\mathcal{F}_q \longleftarrow \Theta(\mathcal{X}_q), \mathcal{F}_s \longleftarrow \Theta(\mathcal{X}_s)$
  $\mathcal{L}_T \longleftarrow \mathcal{L}_{fs} \longleftarrow -\log p(y = k|x_q)$
  Update: $\Phi = \Phi - \alpha\nabla\mathcal{L}_T$
  Evaluate $\hat{\Theta}$ on $\mathcal{Q} \longrightarrow \mathcal{Q}_{seen} \bigcup \mathcal{Q}_{unseen}$;
end

$$\mathcal{L}_T = \mathcal{L}_{fs} + \mathcal{L}_{v2s} \tag{1}$$

- Visual-to-semantic mapping module (Γ): The Visual-to-semantic mapping module, Γ,

aims at learning a mapping from the visual to the semantic space aided by the Semantic Guidance module during the pretraining stage. In the Semantic Guidance module, we employ a set of fully connected layers similar to the visual branches (θ_s/θ_q) to map the vector output from the fastText [10] module to the shared semantic space. Here we employ the simple MSE loss to train the Visual-to-semantic module as shown in Equation 2.

$$\mathcal{L}_{v2s} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\|\Psi(\mathcal{Y}_s) - \Gamma(\mathcal{F}_s)\|_2^2 \tag{2}$$

- Few-shot learning and feature fusion: The output $\mathcal{W}_s$ obtained from the Visual-to-semantic module is fused with the corresponding support visual feature f_s to obtain the semantically augmented feature vector $\mathcal{A}_s$. The query features and the augmented support feature vectors are passed through the corresponding embedding functions, ϕ_q and ϕ_s. We compute a D dimensional prototype vector for each class through the embedding function $\phi_s : \mathbf{R}^M \longrightarrow \mathbf{R}^D$. Each class prototype, μ_k for a class k such that $k \in 1 \ldots N$ in a given episode is computed as the mean vector of the embedded support points as shown in Equation 3, where $\mathcal{A}_{s,j}$ is a feature from the augmented support feature set $\mathcal{A}_s^k$ of the class k.

$$\mu_k = \frac{1}{|\mathcal{A}_s^k|}\sum_{\mathcal{A}_{s,j}\in\mathcal{A}_s^k}\phi_s(\mathcal{A}_{s,j}) \tag{3}$$

Learning progresses by minimizing the objective function in the embedding space over the features $\phi_q(f_q^j)$ corresponding to the query samples, $x_q^j \in \mathcal{X}_q$ for a sample indexed by j as shown in Equation 4,

$$\mathcal{L}_{fs} = -\log p(y = k|x_q) \tag{4}$$

where $p(y = k|x_q)$ for j^{th} sample is defined as,

$$p(y = k|x_q^j) = \frac{exp(-\left\|\mu_k - \phi_q(f_q^j)\right\|_2)}{\sum_l exp(-\left\|\mu_l - \phi_q(f_q^j)\right\|_2)} \tag{5}$$

Simultaneous to the N way K shot learning in the initial stage of training over the base dataset $\mathcal{D}_{base}$, we also train a conditional GAN to generate pseudo samples from the base dataset as in the incremental stage we do not retain any samples from it. Instead, we only need to retain the GAN generator parameters.

- C-GAN for $\mathcal{D}_{base}$ replay: The conditional GAN is trained only using the support set samples of $\mathcal{D}_{base}$ in each episode, with $\mathcal{X}_s$ as the real samples against the fake samples generated from the noise vector z conditioned on the label. We employ a mean square error based adversarial loss function seen in [18] to train the C-GAN as shown in Equation 6 via the standard min-max optimization,

$$\begin{aligned}\min_D \mathcal{J}(D) &= \frac{1}{2}\mathbf{E}_{x\sim p_{data}(\mathcal{X}_s)}[(D(x) - \mathcal{Y}_s)^2] + \frac{1}{2}\mathbf{E}_{x\sim p_z(z)}[(D(G(z)) - y)^2] \\ \min_G \mathcal{J}(G) &= \frac{1}{2}\mathbf{E}_{x\sim p_z(z)}[(D(G(z)) - c)^2]\end{aligned} \tag{6}$$

where y and $\mathcal{Y}_s$ denotes the labels corresponding to fake and real data while c is the label which the generator wants the discriminator to believe for fake samples. While, $p_{data}(\mathcal{X}_s)$

denotes the distribution over the data $\mathcal{X}_s$ and $p_z(z)$ denotes the normal distribution from which the noise vector z is sampled.

b. Incremental learning stage: The incremental stage commences with the introduction of the unseen incoming dataset $\mathcal{D}_{novel}$. In this stage, we freeze weights of all modules except that of the sub-networks corresponding to the embedding functions ϕ_s and ϕ_q as seen in Figure 2. We remove the Semantic Guidance module altogether in this stage, and the semantic mapping module, Γ, operates independently, relying only on the mappings learned from the pre-training stage. The training over the novel dataset commences in a similar episodic fashion as carried out in the pre-training stage.

We sample the support set from the novel dataset and generated pseudo samples of the base classes, and ensure that the corresponding joint query set classes are such that they form a subset of both the base and novel classes. Both these sets are then passed through the encoder Θ to get the features $\mathcal{F}_q$ and $\mathcal{F}_s$, respectively. The features $\mathcal{F}_s$ are then fed to the trained visual to semantic mapping module Γ trained in the pre-training stage to obtain the vector $\mathcal{W}_s$. The vector $\mathcal{W}_s$ concatenated with the corresponding visual support features, f_s will form the augmented support feature, $\mathcal{A}_s$ which is the input to the embedding function ϕ_s. The output of the embedding function ϕ_q corresponding to the unlabelled query set is also obtained over which the model is fine-tuned and evaluated.

Even though we do not retain any samples from the base dataset after the pre-training stage, we need to ensure that the model does not forget previously acquired knowledge. This is done by replaying the base class pseudo samples obtained using the C-GAN trained in the previous stage. We use the generated base samples and the novel unseen samples in equal proportion in every iteration. During the N way K shot incremental training stage, we use only the objective function shown in Equation 4 to tune the parameters of the sub-networks ϕ_s and ϕ_q. As discussed, we evaluate the model simultaneously on the query set from both the base and novel classes again in an N way K shot episodic fashion.

4 Experiments

4.1 Datasets

We consider a total of four standard datasets to evaluate the proposed SemGIF framework: CIFAR100 [3], miniImagenet [39], CUB200-2011 [40] and TieredImagenet [28]. Thus spanning a comprehensive set of datasets, including the two challenging subsets of Imagenet [29]. For each of these datasets, we utilize the train split as $\mathcal{D}_{base}$ and the test split as $\mathcal{D}_{novel}$ which is more challenging due to the absence of the meta-training stage.

4.2 Model Architecture and Implementation Details

We use a Resnet-18 encoder as the backbone feature extractor network to obtain the visual features from the input images. While the image branch and text branch sub-networks used, as seen in Figure 1 are each a two-layer fully connected network. The Visual-to-semantic module is a simple encoder-decoder network using just two fully connected layers each. Similarly, each of the embedding functions uses a minimal three-layer densely connected network. Throughout this implementation, we have used leaky-ReLU as the activation function with a negative slope value of 0.01. We also have employed batch-norm between the layers in the image branch, text branch, and the Visual-to-semantic mapping networks, while the embedding function also utilizes a dropout with $p = 0.5$ between its layers. fastText word

Dataset	CIFAR100		miniImagenet		CUB200		TieredImagenet*	
Method	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**
ProtoNet [39] $\mathcal{B}_1$	40.96%	62.50%	41.07%	55.15%	29.45%	46.00%	30.04%	41.38%
CADA-VAE [31] $\mathcal{B}_2$	-	-	-	-	54.15%	62.05%	-	-
aCASTLE [10] $\mathcal{B}_2$	-	-	43.63 %	56.33%	-	-	22.23%	33.54%
LwoF [9]	-	-	52.37%	59.89%	-	-	52.40%	62.63%
Imprint [25]	-	-	41.25%	43.92%	47.62%	61.59%	39.13%	53.60%
Attractor [27]	-	-	53.62%	62.83%	-	-	56.11%	65.52%
XtarNet [45]	-	-	55.28%	66.86%	-	-	**61.37%**	69.58%
Ours without semantic	42.86%	55.17%	38.98%	49.89%	39.74%	63.29%	49.65%	62.83%
Ours (full)	**57.21%**	**79.95%**	**56.75%**	**75.49%**	**57.35%**	**86.56%**	55.63%	**69.89%**

Table 1: A comparative study with existing algorithms in the literature. We report the Harmonic mean (Acc_H) obtained for our framework across the four datasets under consideration. ('-' denotes the value is not reported). * For TieredImagenet performance reported by our model shows lesser improvement as our approach does not rely on a meta-learning stage despite which we beat the SOTA on other datasets with a considerable margin.

embedding trained on Wikipedia is used in the semantic guidance module, which is active during the pretraining stage. C-GAN architecture used in this work is adopted from the implementation available here[1]. Throughout the training process in both stage, we use the Adam optimiser [13] with a learning rate of 0.0001 and weight decay of 0.0001. C-GAN's generator and the discriminator are trained during the initial stage with a learning rate of 0.0002 which again uses the Adam optimizer with the beta values set as 0.5 and 0.999. We implemented our model using a single Nvidia GeForce GTX 1080 Ti GPU.

Evaluation Protocols: For all the results discussed in this paper, we have used classification accuracy to evaluate the performance.

In the incremental learning stage for evaluating the model, we sample a few-shot episode consisting of classes sampled from both base and novel classes. The model accuracy is then reported on the joint evaluation over the query set with classes consisting of the seen base classes and unseen novel classes. We have maintained an equal proportion of base and novel classes in our experiments. For the experiments, say for N-way K-shot, we consider K support examples. (i.e., shots) Furthermore, we use a query set of 15 samples per class from base and novel datasets. We also report the individual accuracy of the model over both the unseen and seen data during the incremental stage as shown in Table 3. We use pretrain Acc_b to denote the accuracy over the base classes during the initial stage, while Inc Acc_b denotes the same achieved during the incremental stage over the generated base image samples. While Acc_n is used to report the model's accuracy over the novel dataset and Acc_H denotes the harmonic mean. For the incremental stage where evaluation over the base and novel dataset is considered, we report the harmonic mean as defined in [42]. For this we calculate the accuracies Acc_b and Acc_n separately to compute the harmonic mean, Acc_H.

Table 1 shows the comparison of performance of the proposed framework with the existing literature in both the incremental few-shot [9, 25, 27, 34, 45] and generalised few-shot domains [32, 44]. The traditional FSL work in [34] acts as our preliminary baseline $\mathcal{B}_1$, while the GFSL works in [32, 44] will serve as the baseline comparison $\mathcal{B}_2$ for the performance over the combined evaluation over the base and novel dataset with the data from the base dataset still being accessible. While the IFSL works in [9, 25, 27, 45]are used as the final set of baseline. Note that unlike other algorithms in the literature, we evaluate our frame-

[1] https://github.com/eriklindernoren/PyTorch-GAN

Dataset	CIFAR100		miniImagenet		CUB200		TieredImagenet	
	5way-5shot	2way-5shot	5way-5shot	2way-5shot	5way-5shot	2way-5shot	5way-5shot	2way-5shot
Pretrain	74.93%	95.61%	69.00%	95.54%	90.00%	89.27%	46.00%	66.70%
Incremental	88.57%	96.77%	83.57%	95.56%	87.00%	86.98%	69.58%	78.68%

Table 2: 5 shot Results for SemGIF using Resnet-50 backbone encoder

Dataset	CIFAR100				miniImagenet				CUB200				TieredImagenet			
	pretrain Acc_b	Inc Acc_H	Inc Acc_b	Inc Acc_n	pretrain Acc_b	Inc Acc_H	Inc Acc_b	Inc Acc_n	pretrain Acc_b	Inc Acc_H	Inc Acc_b	Inc Acc_n	pretrain Acc_b	Inc Acc_H	Inc Acc_b	Inc Acc_n
5-way 5-shot	66.52	79.95	87.27	73.76	63.63	75.49	86.09	67.22	84.41	86.56	91.09	82.46	47.23	69.89	83.16	60.27
2-way 5-shot	93.40	91.23	84.82	98.68	95.41	95.21	92.44	98.15	88.12	83.56	88.22	79.37	69.44	79.71	94.39	68.98
5-way 1-shot	44.22	57.21	75.50	46.06	40.22	56.75	71.89	46.87	43.01	57.35	67.31	49.96	33.04	55.63	67.25	47.43
2-way 1-shot	69.08	75.74	74.46	77.06	65.21	75.16	71.32	79.44	64.13	58.44	70.30	50.00	60.03	62.23	71.19	55.28
5-way 10-shot	72.65	87.61	94.64	81.54	69.82	83.39	92.09	76.20	91.56	87.45	85.00	90.03	51.85	68.47	94.20	53.78
2-way 10-shot	96.24	97.74	96.67	98.84	96.27	96.59	94.52	98.75	90.00	89.24	88.22	79.37	71.55	81.95	95.64	71.68

Table 3: IFSL Results on multiple combinations of N-way K-shot across the four datasets

work across a wider variety of datasets. We also evaluate the performance in a cross-domain setting where $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ are derived from completely different datasets as in 4.

4.3 Results

From the results shown in Table 1, it is clear that the proposed semantic augmentation-based approach outperforms the existing algorithms[2]. For the miniImagenet dataset, our framework outperforms all other methods and achieves a relative improvement of 8.63% in 5-way 5-shot and 1.47% in 5-way 1-shot over the closest performing method [45]. Similarly, our approach shows an increase in performance on CUB-200 of about 24.51% in 5-way 5-shot and 3.2% in the 5-way 1-shot case in comparison with [32]. Tieredimagenet is a larger subset of the ILSVRC-12, which is more realistic and challenging as argued in [27]. For this, our approach yields superior performance in the 5-way 5-shot setting, with a relative improvement of 0.31% over the closest performing method [45], while our 5-way 1-shot result is only marginally inferior. This dip in performance can mainly be attributed to the high variability in the test class samples relative to the base classes learned. Meanwhile, our model shows an improvement in performance on CIFAR-100 on both the 5-way 5-shot and 5-way 1-shot classification by 17.45% and 16.25% respectively, with [34] being the only available approach to compare to.

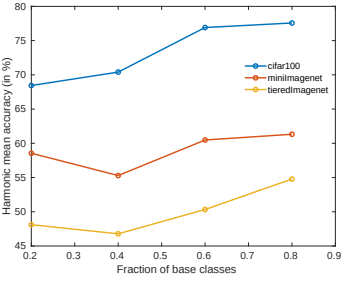


Figure 3: Change in accuracy with the variation in the number of base classes considered

4.4 Ablation study

Multiple combinations of N-way K-shot: Table 3 shows more experimental results using multiple combinations of N-way K-shot in a detailed manner. It can be observed that the proposed method shows significantly less forgetting of the base classes and even depict improved performance on base classes during the incremental stage across all the datasets. The superior discriminative nature of our fine-grained features to which the performance of our model can be attributed is established by the t-SNE plots for both the base and novel dataset shown in Figure 4 for both the miniImagenet and TieredImagenet datasets.

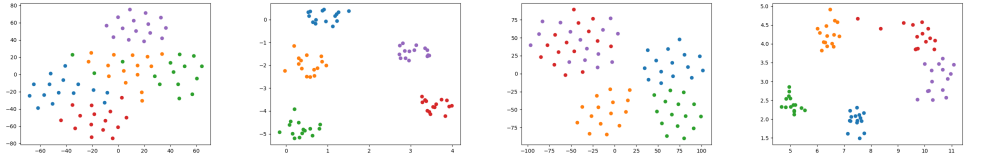
[2]More quantitative and qualitative results are included in the supplementary material.

Base Dataset	CIFAR100		miniImagenet		CUB200		TieredImagenet	
Novel Dataset	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**	**5way-1shot**	**5way-5shot**
CIFAR100	-	-	60.73%	81.00%	57.91%	78.63%	54.87%	78.45%
miniImagenet	55.89%	75.95%	-	-	51.41%	75.68%	50.69%	74.86%
CUB200	56.32%	84.61%	57.19%	83.03%	-	-	50.38%	82.19%
TieredImagenet	55.19%	69.07%	55.88%	67.75%	51.42%	69.49%	-	-

Table 4: Heterogeneous evaluation by choosing $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ from different domains

Heterogeneous evaluation study: We perform a heterogeneous evaluation study using the standard datasets considered. That is, we create the base dataset $\mathcal{D}_{base}$ from one dataset while the novel dataset $\mathcal{D}_{novel}$ in the incremental learning stage comes from a different domain. We observe that the proposed method shows excellent performance consistently across the datasets even when there is a domain shift between the $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$. The results for this experiment are shown in Table 4.

Effect of the number of base classes: We also study the effect of the number of training classes used during the pre-training stage. This, as expected, shows an upward trend for accuracy with the increase in the number of base classes considered as seen in Figure 3.



(a) miniImagenet base (b) miniImagenet novel (c) TieredImagenet base (d) TieredImagenet novel

Figure 4: TSNEs for both generated base and incoming novel images from $\mathcal{D}_{base}$ for miniImagenet and Tieredimagenet datasets. Each color indicates a different class.

Effect of encoder size: Finally, Table 2 shows how the model performance varies with the increase in the size of the visual encoder Θ used. Although we observe a performance gain with the increase of capacity of the feature extractor as expected, the extent of this boost is not in proportion with the growth of visual encoder from resnet-18 to resnet-50. Thus it is safe to assume that even though the role played by the visual encoder is significant in our model; it does not supersede the performance gain resulted from the introduction of our novel SemGIF approach. This notion is again reinforced by the results shown in Table 1, wherein we show how the performance of the model degrades in the absence of semantic fusion.

5 Conclusions

We proposed SemGIF, a generative modeling-based IFSL framework that uses an additional semantic space to generate discriminative class prototypes. The conditional GAN-based generative module helps synthesize data from the base classes on demand, which subsequently helped incorporate the base knowledge while incrementally adapting the model to the novel few-shot categories. We have performed extensive experiments on four benchmark datasets where we consistently observed our model outperform the literature on different IFSL settings. We further introduced a heterogeneous experimental scenario where the base and novel classes originate from distinct datasets. In the future, we are interested in extending our framework to support open-set classes in incremental episodes.

References

[1] Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, and Ethan Fetaya. Gp-tree: A gaussian process classifier for few-shot incremental learning. *arXiv preprint arXiv:2102.07868*, 2021.

[2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip HS Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *arXiv preprint arXiv:1606.05233*, 2016.

[3] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[5] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[7] Michael Fink. Object classification from a single example utilizing class relevance metrics. *Advances in neural information processing systems*, 17:449–456, 2005.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[9] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[11] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017.

[12] Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, and In So Kweon. Variational prototyping-encoder: One-shot learning with prototypical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2019.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[15] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2021.

[16] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3287, 2017.

[17] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[18] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[19] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. *arXiv preprint arXiv:2103.00991*, 2021.

[20] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[21] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[22] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

[23] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11321–11329, 2019.

[24] Tomas Pfister, James Charles, and Andrew Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *European Conference on Computer Vision*, pages 814–829. Springer, 2014.

[25] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.

[26] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[27] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard S Zemel. Incremental few-shot learning with attention attractor networks. *arXiv preprint arXiv:1810.07218*, 2018.

[28] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[30] Ruslan Salakhutdinov, Joshua Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 195–206. JMLR Workshop and Conference Proceedings, 2012.

[31] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

[32] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[33] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017.

[34] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[35] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[36] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020.

[37] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646. MORGAN KAUFMANN PUBLISHERS, 1996.

[38] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. *arXiv preprint arXiv:1707.02610*, 2017.

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

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

[41] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.

[42] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[43] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6619–6628, 2019.

[44] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. Learning adaptive classifiers synthesis for generalized few-shot learning. *International Journal of Computer Vision*, pages 1–24, 2021.

[45] Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning. In *International Conference on Machine Learning*, pages 10852–10860. PMLR, 2020.

[46] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2021.