This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

OrCo: Towards Better Generalization via Orthogonality and Contrast for Few-Shot Class-Incremental Learning

Noor Ahmed* Anna Kukleva* Bernt Schiele

{noahmed, akukleva, schiele}@mpi-inf.mpg.de

Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

Few-Shot Class-Incremental Learning (FSCIL) introduces a paradigm in which the problem space expands with limited data. FSCIL methods inherently face the challenge of catastrophic forgetting as data arrives incrementally, making models susceptible to overwriting previously acquired knowledge. Moreover, given the scarcity of labeled samples available at any given time, models may be prone to overfitting and find it challenging to strike a balance between extensive pretraining and the limited incremental data. To address these challenges, we propose the OrCo framework built on two core principles: features' orthogonality in the representation space, and contrastive learning. In particular, we improve the generalization of the embedding space by employing a combination of supervised and self-supervised contrastive losses during the pretraining phase. Additionally, we introduce OrCo loss to address challenges arising from data limitations during incremental sessions. Through feature space perturbations and orthogonality between classes, the OrCo loss maximizes margins and reserves space for the following incremental data. This, in turn, ensures the accommodation of incoming classes in the feature space without compromising previously acquired knowledge. Our experimental results showcase state-of-the-art performance across three benchmark datasets, including mini-ImageNet, CIFAR100, and CUB datasets. Code is available at: https://github.com/noorahmedds/OrCo.

1. Introduction

Real-world applications frequently encounter various challenges when acquiring data incrementally, with new information arriving in continuous portions. This scenario is commonly referred to as Class Incremental Learning (CIL) [2–4, 13, 14, 24, 25, 31, 33, 39, 43, 48]. Within CIL, the foremost challenge lies in preventing catastrophic forgetting [10, 19, 27], where previously learned concepts are susceptible to being overwritten by the latest updates. However, in a Few-Shot Class-Incremental Learning (FSCIL) scenario [8, 12, 12, 18, 22, 26, 35, 40, 42, 44–46], characterized by the introduction of new information with only a few labeled samples, two additional challenges emerge: overfitting and intransigence [7, 34]. Overfitting arises as the model may memorize scarce input data and lose its generalization ability. On the other hand, intransigence involves maintaining a delicate balance, preserving knowledge from abundant existing classes while remaining adaptive enough to learn new tasks from a highly limited dataset. Advances in dealing with catastrophic forgetting, overfitting and intransigence are important steps toward improving the practical value of these methods.

Catastrophic forgetting is commonly tackled in CIL methods [14, 16, 31], which assume ample labeled training data. However, standard CIL methods struggle in scenarios with limited labeled data, such as FSCIL [35]. To

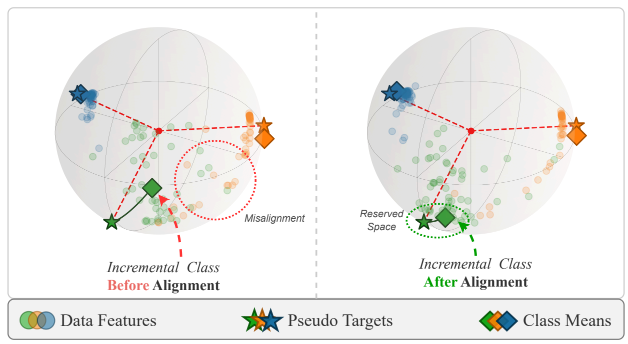

Figure 1. **PCA analysis on feature space before and after alignment.** Left: Before aligning incremental classes to orthogonal pseudo-targets. Right: After aligning incremental classes to assigned targets using **OrCo** loss. Our loss effectively reduces misalignment. Additionally, it enhances generalization for incoming classes by explicitly reserving space.

* Equal Contribution

address the three challenges posed by FSCIL, recent approaches [42, 45] focus primarily on regularizing the feature space during incremental sessions, mitigating the risk of overfitting. These methods rely on a frozen backbone pretrained with standard cross-entropy on a substantial amount of data from the base session. However, we argue that achieving high performance on the pretraining dataset may not necessarily result in optimal generalization in subsequent incremental sessions with limited data. Therefore, in the first phase, we propose enhancing feature space generalization through contrastive learning, leveraging data from the base session.

In this work, we introduce the OrCo framework, a novel approach built on two fundamental pillars: features' mutual orthogonality on the representation hypersphere and contrastive learning. During the first phase, we leverage supervised [17] and self-supervised contrastive learning [6, 11, 29] for pretraining the model. The interplay between these two learning paradigms enables the model to capture various types of semantic information that is particularly beneficial for the novel classes with limited data [1, 5, 15], implicitly addressing the *intransigence* issue. After the pretraining, we generate and fix mutually orthogonal random vectors, further referred to as pseudo-targets. In the second phase, we aim to align the fixed pretrained backbone to the pseudo-targets using abundant base data. The learning objective during this phase is our OrCo loss, which consists of three integral components: perturbed supervised contrastive loss (PSCL), loss term that enforces orthogonality of features in the embedding space, and standard cross-entropy loss. Notably, our PSCL leverages generated pseudo-targets to maximize margins between the classes and to preserve space for incremental data, enhancing orthogonality through contrastive learning (see figure 1). The third phase of our framework, which we apply in each subsequent incremental session, similarly aims to align the model with the pseudo-targets, but using only few-shot data from the incremental sessions. During the third phase, our PSCL addresses limited data challenges, mitigating the *overfitting* problem to the current incremental session and *catastrophic forgetting* of the previous sessions through margin maximization.

We summarize the contributions of this work as follows:

- We introduce the novel OrCo framework designed to tackle FSCIL, that is built on orthogonality and contrastive learning principles throughout both pretraining and incremental sessions.
- Our perturbed supervised contrastive loss introduces perturbations of orthogonal, data-independent vectors in the representation space. This approach induces increased margins between classes, enhancing generalization.
- We showcase robust performance on three datasets, outperforming previous state-of-the-art methods. Furthermore, we perform a thorough analysis to evaluate the importance of each component.

2. Related Work

Few-Shot Learning (FSL). In FSL, a model is trained on scarce data with just a few samples per class. Current literature can be divided into two predominant categories: optimisation-based [9, 28, 30] and metric-based methods [30, 34, 37]. Optimization-based approaches, such as MAML [9] and Reptile [28], find optimal parameters that can generalize quickly on other sets when subjected to fine-tuning. Conversely, metric-based methods utilize a pretrained model and compare support and query instances using similarity metrics. For example, Prototypical Networks [34] learns a metric space where the distance to class prototypes determine classification. And imprinting weights method [30] shows improved performance by using class means as strong initialisation for an evolving classifier and integrates principles from both categories.

Class-Incremental Learning (CIL). In the domain of CIL, a sequence of novel concepts must be learned without forgetting previously acquired knowledge. Recent works can be coarsely categorized in 3 groups. Foremost, there are knowledge distillation schemes [13, 24, 31, 39], which retain model behaviour across the adaptation process to avoid forgetting. Then, data-replay methods [2, 3, 14, 43, 48] show strong resistance to catastrophic forgetting by storing old class exemplars. Lastly, weight consolidation methods [4, 19, 25, 33] identify important weights and moderate training regimen.

Few-Shot Class-Incremental Learning (FSCIL). FSCIL paradigm demands rapid adaptation to novel classes with limited data. The methods can be divided into the following categories [36]: geometry preservation methods [35, 42], replay or distillation strategies [22, 45], metric learning methods [8, 18, 44, 46] and meta-learning [12, 26, 40], highlighting the breadth of methodologies in FSCIL. Parallel to our methodology, metric learning methods utilize tricks in the feature space, showcasing diverse approaches for accommodating incremental classes. FACT [46] creates virtual prototypes to reserve space and scale the model for incoming classes. NC-FSCIL [40], aligns class features with the classifier prototypes, which are formed as a simplex equiangular tight frame, using dot-regression loss. C-FSCIL [12] aligns class prototypes quasi-orthogonally to negate interference between classes. In contrast, our approach stands out for its use of contrastive learning with data agnostic pseudo-targets and margin maximization through perturbations in the embedding space improving generalization in incremental sessions.

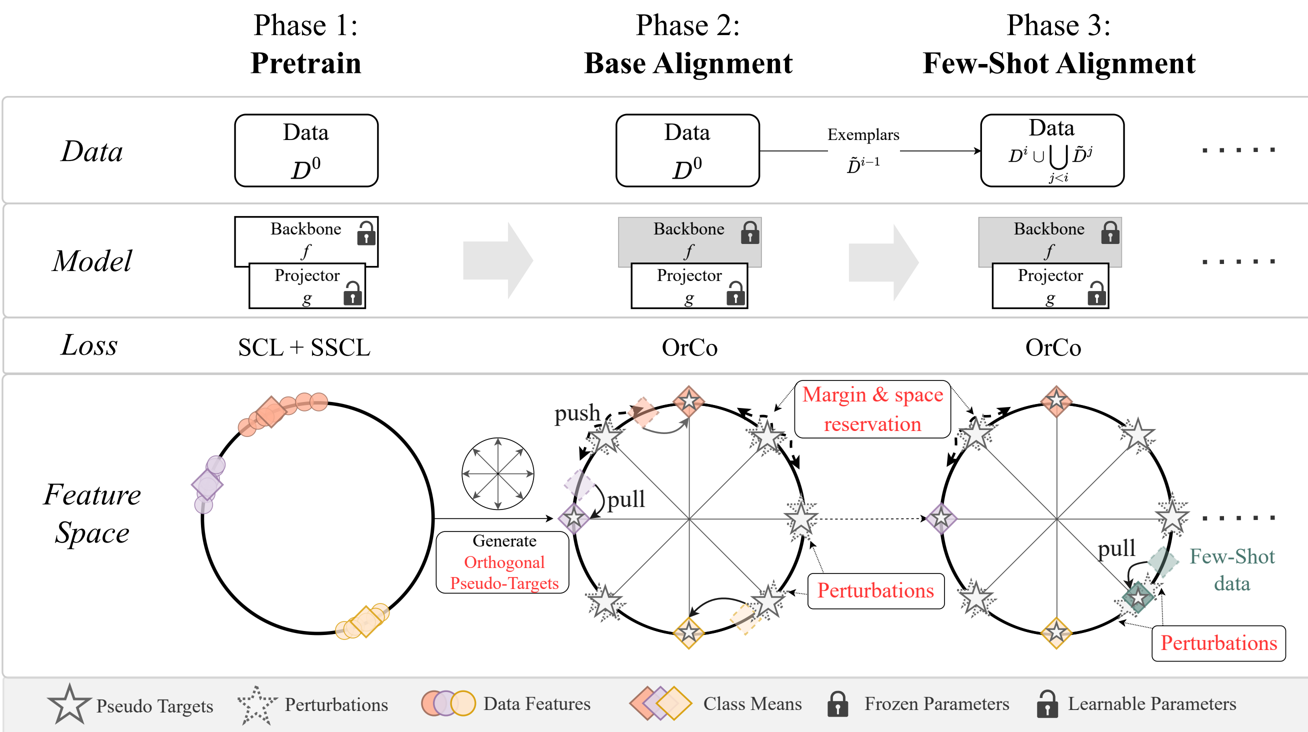

Figure 2. **Overview of OrCo framework.** Our OrCo framework is a three-phase approach for FSCIL. Phase 1 (Pretrain): We pretrain both backbone and projection head with SCL and SSCL on base dataset D^0. Before the next phase, we generate mutually orthogonal pseudo-targets. Phase 2 (Base Alignment): We aim to align the base dataset D^0 to the pseudo-targets through our OrCo loss. This involves pulling class features towards the nearest pseudo-targets and pushing forces based on perturbations around unassigned pseudo-targets (grey stars without assigned colored class means) to increase the margin and preserve space for incoming classes. Phase 3 (Few-Shot Alignment): Phase 3, employed in each subsequent incremental session, is similar to Phase 2 and assigns pseudo-targets to incremental class means with further alignment using our OrCo loss.

3. OrCo Framework

We begin with necessary preliminaries in section 3.1, followed by the description of our OrCo framework in section 3.2 and OrCo loss in section 3.3.

3.1. Preliminaries

FSCIL Setting. FSCIL consists of multiple incremental sessions. An initial 0-th session is often reserved to learn a generalisable representation on an abundant base dataset. This is followed by multiple few-shot incremental sessions with limited data. To formalise, an M-Session N-Way and K-Shot FSCIL task consists of $D_{seq} = \{D^0, D^1, ..., D^M\}$. These are all the datasets written in sequence where $D^i = \{(x_i, y_i)\}_{i=1}^{|D^i|}$ is the dataset for the i-th session. The 0-th session dataset D^0, also referred to as base dataset, consists of C^0 classes, each with a large number of samples. The training set for each following few-shot incremental session ($i > 0$) has N classes. Each of these classes has K samples, typically ranging from 1 to 5 samples per class. Taking the i-th session as an example, the model's performance is assessed on validation sets from the current (i-th) and all previously encountered datasets ($< i$). The entire FSCIL task comprises a total of C classes. In our OrCo framework, we use base dataset D^0 for pretraining the model during phase 1 and for base alignment during phase 2.

Target Generation. We employ a Target Generation loss, similarly as in [23], to generate mutually orthogonal vectors across the representation hypersphere with a dimensionality of d. First, we define a set of random vectors $T = \{t_i\}$ where $\{t_i\} \in \mathbb{R}^d$. The optimization of the following loss with respect to these random vectors maximizes the angle between any pair of vectors $t_i, t_j \in T$, thereby ensuring their mutual orthogonality:

$$\mathcal{L}_{TG}(T) = \frac{1}{|T|} \sum_{i=1}^{|T|} \log \sum_{j=1}^{|T|} e^{t_i \cdot t_j / \tau_o} \tag{1}$$

where τ_o is the temperature parameter. These optimized vectors, which we further refer to as pseudo-targets, remain fixed throughout our training process.

Contrastive Loss. The objective of contrastive representation learning is to create an embedding space where similar sample pairs are in close proximity, while dissimilar pairs are distant. In this work, we adopt the InfoNCE loss [17, 29] as our contrastive objective. With positive set P_i and negative set N_i defined for each data sample z_i, called anchor, this loss aims to bring any $z_j \in P_i$ closer to its anchor z_i and push any $z_k \in N_i$ further away from the anchor z_i:

$$\mathcal{L}_{CL}(i;\theta) = \frac{-1}{|P_i|} \sum_{z_j \in P_i} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum\limits_{z_k \in N_i} \exp(z_i \cdot z_k/\tau)}, \quad (2)$$

where τ is the temperature parameter. In the classical self-supervised contrastive learning (SSCL) scenario, where labels for individual instances are unavailable, the positive set comprises augmentations of the anchor, and all other instances are treated as part of the negative set [29]. In contrast, for the supervised contrastive loss (SCL), the positive set includes all instances from the same class as the anchor, while the negative set encompasses instances from all other classes [17]. To enhance clarity, we denote the supervised contrastive loss as $\mathcal{L}_{SCL}$ and self-supervised contrastive loss as $\mathcal{L}_{SSCL}$. We consider SCL and SSCL as the cornerstone guiding our work due to their discriminative nature, robustness, and extendability. We leave formal definition of the losses for the supplement.

3.2. OrCo Framework

Overview. Our OrCo framework (see figure 2) for FSCIL begins with a pretraining of the model in the first phase, focusing on learning representations, which are transferable to the new tasks. To achieve this, we leverage both supervised and self-supervised contrastive losses [5, 15]. Before the second phase, we generate a set of mutually orthogonal vectors, which we term as pseudo-targets. In the subsequent second phase, referred to as base alignment in figure 2, we allocate pseudo-targets to class means and ensure alignment through our OrCo loss, using abundant base data D^0. The third phase, implemented in each subsequent incremental session, similarly focuses on assigning incoming but few-shot data to unassigned pseudo-targets, followed by alignment through our OrCo loss. Our OrCo loss comprises three key components: cross-entropy, orthogonality loss, and our novel perturbed supervised contrastive loss (PSCL). The cross-entropy loss aligns incremental data with assigned fixed orthogonal pseudo-targets, the orthogonality loss enforces a geometric constraint on the entire feature space to mimic the pseudo-targets distribution, and our PSCL enhances crucial robustness for FSCIL tasks through margin maximization and space reservation, leveraging mutual orthogonality of pseudo-targets.

Phase 1: Pretrain. In the first pretraining phase, we learn an encoder that accumulates knowledge and generates distinctive features. Using a combination of supervised contrastive loss (SCL) and self-supervised contrastive loss (SSCL), we enhance feature separation within classes, improving model transferability to incremental sessions [5, 15]. To this end, we train the model encoder f and MLP projection head g using base data D^0, mapping input images to $\mathcal{R}^d$ feature space. The pretraining loss is then defined as:

$$\mathcal{L}_{pretrain}(D^0; f, g) = (1-\alpha) * \mathcal{L}_{SCL} + \alpha * \mathcal{L}_{SSCL}, \quad (3)$$

where α controls the contribution of each contrastive loss.

Pseudo-targets. During the first phase, we do not employ any explicit class vectors that can be used for linear classification. Therefore, we generate data-independent mutually orthogonal pseudo-targets $T = \{t_j\} \in \mathcal{R}^d$ on the hypersphere by optimizing loss shown in equation 1, where $|T| >= C$. Further, these pseudo-targets are fixed and assigned to classes, which, in turn, maximize margins between the classes and improve generalization.

Phase 2: Base Alignment. In addition to the pretraining phase, we introduce the second phase based on the base dataset D^0. This phase initiates alignment between the projection head g and the set of generated pseudo-targets T. More specifically, we create class means by averaging features with the same labels. Then, we employ a one-to-one matching approach, utilizing the Hungarian algorithm [21], to assign class means with the most fitting set of pseudo-targets T^0, where $|T^0| = |C^0|$. Note that each class $y_j \in C^0$ is then associated with the respective pseudo-target $t_j^0 \in T^0$ and we denote the remaining unassigned pseudo-targets as $T_u^0 = T \setminus T^0$. Further, we use pseudo-targets T^0 as base class representations for classification. Despite the optimal initial assignment, we enhance the alignment of the projection head g and the respective pseudo-targets T^0 through the optimization of our OrCo loss. Further insights into the specifics and motivation behind our OrCo loss, we elaborate on in section 3.3.

Phase 3: Few-Shot Alignment. In the third phase of our framework, applied in each subsequent incremental session, our goal remains to align incoming data with the pseudo-targets. The following sessions introduce few shot incremental data D^i for the i-th session. Building on previous methods [3, 8, 22, 39], we maintain some exemplars from previously seen classes, constituting a joint set $D^i_{joint} = \{D^i \cup \{\bigcup_{j=0}^{i-1} \tilde{D}^j\}\}$, where $\tilde{D}^j$ denotes saved exemplars from earlier sessions. Keeping random exemplars from previous sessions serves to mitigate both overfitting and catastrophic forgetting issues. Similarly to the second phase, we assign pseudo-targets to incremental class means. To be specific, we determine the optimal assignment between T_u^{i-1} and the current class means, resulting in the optimal

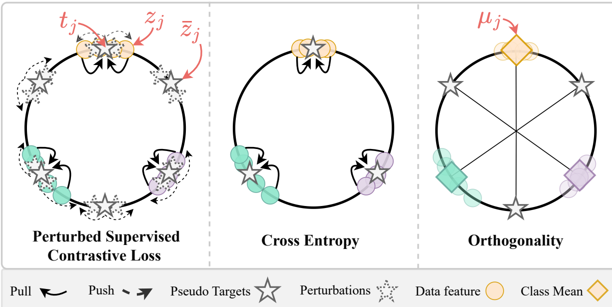

Figure 3. **OrCo loss** consists of three components: our proposed perturbed supervised contrastive loss (PSCL), cross-entropy loss (CE), and orthogonality loss (ORTH). z_j denotes the real data anchor point for a contrastive loss, $\bar{z}_j$ denotes the unassigned pseudo-target anchor, and t_j denotes an additional positive sample for yellow class in the form of an assigned pseudo-target. μ_j represents the within-batch mean features.

assignment set of pseudo-targets T^i. Respectively, the unassigned set of pseudo-targets becomes $T^i_u = T^{i-1}_u \setminus T^i$. During incremental session i, we optimize our OrCo loss given data D^i_{joint}, pseudo-targets $T = \{T^i\}^i_0 \cup T^i_u$ and the assignment between $T = \{T^i\}^i_0$ and the respective classes.

3.3. OrCo Loss

During the second and the third phases, we optimize the parameters of the projection head g with our OrCo loss. This loss comprises three integral components: our novel perturbed supervised contrastive loss (PSCL), cross-entropy loss (CE) and orthogonality loss (ORTH), see figure 3. The aim of optimizing the OrCo loss is to align classes with their assigned pseudo-targets, simultaneously maximizing the margins between classes. This, in turn, enhances overall generalization performance.

To maximize the margins in the representation space, we introduce uniform perturbations of the pseudo-targets T, resulting in perturbed pseudo-targets $\tilde{T} = \{\tilde{t}_j\}^{|T|}_0$ defined as

$$\tilde{t}_j = t_j + \mathcal{U}(-\lambda, \lambda), \tag{4}$$

where $\mathcal{U}$ stands for uniform distribution and λ defines sampling boundaries. To utilize the introduced perturbations, we redefine positive P_j and negative N_j sets for the contrastive loss for the anchor z_j in equation 3. Note that during incremental session i the positive set P^i_j in the standard SCL contains all $z_k \in D^i_{joint}$ such that the label y_k is equal to anchor label y_j, e.g. in figure 3, all yellow circles belong to the positive set for yellow anchor z_j. And the negative set consists of remaining samples $N^i_j = D^i_{joint} \setminus P^i_j$, in figure 3, the negative set is composed of all other colors.

To adapt standard SCL to PSCL, we expand the definition of the positive set. The anchor z_j, with its assigned pseudo-target $t_j \in T$, becomes an additional positive pair, see figure 3. Furthermore, considering the previously defined pseudo-target perturbations, we incorporate them into the positive set, resulting in $\tilde{P}^i_j = P^i_j \bigcup t_j \bigcup \tilde{t}_j$. In figure 3, the positive set for yellow anchor z_j contains all yellow circles and additionally the pseudo-target t_j with its perturbations. This extension of the positive set introduces additional pushing forces for incremental classes and, therefore, enables the maximization of margins between classes. We show that this approach proves to be especially advantageous in scenarios with limited samples, as it mimics augmentations in the feature space.

On the other hand, we expand the anchor definition. In standard SCL, each anchor z_j belongs to the set of real training data D^i_{joint}, e.g. in figure 3, anchors for standard SCL are only circles. However, we propose to use anchors from both real data and unassigned pseudo-targets (circles and unassigned stars in figure 3), specifically $z_j \in D^i_{joint}$ and $\bar{z}_j \in T^i_u$. The positive set for the anchor $\bar{z}_j \in T^i_u$ (unassigned pseudo-target) contains only corresponding perturbed pseudo-targets $\tilde{P}^i_j = \{\tilde{t}_j\}$ (dashed stars around $\bar{z}_j$ in figure 3), while the negative set $\tilde{N}^i_j = \{D^i_{joint} \bigcup \tilde{T}^i_u\} \setminus \{\tilde{t}_j\}$ includes all real data and other perturbed pseudo-targets. This approach ensures that each unassigned pseudo-target pushes all other classes away, thereby promoting space preservation for the following incremental sessions.

To complement PSCL, we employ cross-entropy loss for sample z_j that pulls class features to their assigned targets during the few-shot incremental session i:

$$\mathcal{L}_{CE}(z_j) = - \sum_{c \in \{C^I\}^i_1} y_c \log \frac{\exp(z_j t_c^T)}{\sum_{k \in \{C^I\}^i_1} \exp(z_k t_c^T)}. \tag{5}$$

We further employ the orthogonality loss ($\mathcal{L}_{ORTH}$) defined similarly as in equation 1. It differs, however, in that it uses the mean class features μ_j (see figure 3) as input and enforces an intrinsic geometric constraint on the feature landscape. See section 5 in supplementary for details.

Finally, our OrCo loss is a combination of the three losses introduced above:

$$\mathcal{L}_{OrCo} = \mathcal{L}_{PSCL} + \mathcal{L}_{CE} + \mathcal{L}_{ORTH}. \tag{6}$$

During testing, a sample is assigned a label based on the nearest assigned pseudo target.

4. Experimental Results

In section 4.1, dataset and evaluation protocol are introduced. Then we compare with state-of-the-art methods on 3 popular benchmarks in section 4.2. Finally, we validate the effectiveness of each of the components in section 4.3.

Method	Base Acc	Session-wise Harmonic Mean (%) ↑								aHM	ΔaHM
		1	2	3	4	5	6	7	8		
IW [30]	83.10	49.49	45.09	45.98	46.30	44.67	42.48	43.26	45.65	45.36	**+12.76**
FACT [46]	75.78	27.20	27.84	27.94	25.17	22.46	20.54	20.88	21.25	24.16	**+33.96**
CEC [42]	72.17	31.91	31.84	30.98	30.74	28.14	26.78	26.96	27.42	29.35	**+28.78**
C-FSCIL [12]	76.60	9.74	20.53	28.68	31.91	34.85	35.05	37.72	37.92	29.55	**+28.57**
LIMIT [47]	73.27	40.34	33.58	31.81	31.74	29.32	29.11	29.57	30.28	31.97	**+26.15**
LCwoF [22]	64.45	41.24	38.96	39.08	38.67	36.75	35.47	34.71	35.02	37.49	**+20.63**
BiDist [45]	74.67	42.42	43.86	43.87	40.34	38.97	38.01	36.85	38.47	40.35	**+17.77**
NC-FSCIL [40]	**84.37**	62.34	61.04	55.93	53.13	49.68	47.08	46.22	45.57	52.62	**+5.50**
OrCo	83.30	**68.71**	**63.87**	**60.94**	**57.98**	**55.27**	**52.41**	**52.68**	**53.12**	**58.12**	

Table 1. **Sota comparison on mini-ImageNet.** aHM denotes the average of the harmonic mean across all sessions. IW [30] is evaluated based on the model learning in our pretrain phase. Detailed results of the individual sessions are in the supplement.

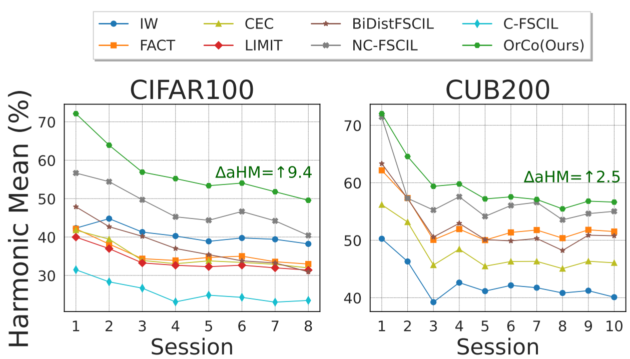

Figure 4. **Sota comparisons on CIFAR100 and CUB200 datasets.** Performance curves, that measure harmonic mean, of our method comparing to recent sota methods. Left: CIFAR100. Right: CUB200. ΔaHM denotes the average harmonic mean improvement over the runner-up method.

4.1. Datasets and Evaluation

We conduct evaluation of our OrCo framework on three FSCIL benchmark datasets: mini-ImageNet [32], CIFAR100 [20] and CUB200 [38]. In the setting formalised by [35] mini-ImageNet and CIFAR100 are organized into 60 base classes and 40 incremental classes structured in a 5-way, 5-shot FSCIL scenario for a total of 8 sessions. CUB200, a dataset for fine-grained bird species classification, contains equal number of base and incremental classes for a total of 200 classes. The dataset is organised as a 10-way, 5-shot FSCIL task and presents a rigorous challenge.

Performance measure Commonly used FSCIL datasets all have a quantity bias towards the base classes. mini-ImageNet and CIFAR100 have both 60% of the data in the base classes and CUB200 with 50%. Consequently, standard accuracy measures like Top-1 accuracy will be skewed in favour of the base-classes. For instance, a method which has a base accuracy $A_{base} = 100\%$ on CUB200 and performs weakly on the first incremental session $A^1_{inc} = 10\%$ would produce a Top-1 average accuracy $A^1_{cls} = 91.82\%$. At first glance, this accuracy may not entirely represent inherent biases in a method, though such measures are commonly used to benchmark performance. To tackle this, harmonic mean has risen as a robust evaluation measure in FSCIL [22, 45, 46]. In the given scenario, the harmonic mean would penalise the method aggressively resulting in a metric score of $A^1_{hm} = 18.18\%$, accurately indicating bias. More concretely, we compute harmonic mean by combining base class accuracy and incremental session accuracy: $A^j_{hm} = (2 \times A_{base} \times A^j_{inc})/(A_{base} + A^j_{inc})$. In addition to this, we propose average harmonic mean (aHM) which is simply averaging the harmonic mean scores from all sessions for a consolidated view.

Implementation Details Our model is optimised using LARS [41] for the pretraining phase and SGD with momentum for phase 2 and 3. For CUB200 dataset, we skip the pretraining following [35, 40, 42] and initialize the model with ImageNet pretrained weights. For the second and third phase, we finetune only the projection head. For the PSCL loss we choose a perturbation magnitude λ = 1e-2. We train the projection head for 10 epochs during the second phase and 100 epochs for the third phase. Cosine scheduling is employed with a maximum learning rate set to 0.1. Augmentations include, random crop, random horizontal flip, random grayscale and a random application of color jitter. Details can be found in the supplement section 6.

4.2. Comparison to state-of-the-art

In this section, we conduct a comparative analysis of our proposed OrCo with recent state-of-the-art approaches. Ta-

PSCL	CE	ORTH	Session-wise Harmonic Mean (%) ↑								aHM
			1	2	3	4	5	6	7	8	
	✓		65.46	56.29	44.12	36.96	26.90	21.11	18.90	16.19	35.74
	✓	✓	65.30	56.21	43.96	37.30	28.31	22.01	19.66	16.64	36.17
✓			50.70	45.42	42.68	39.84	38.71	37.94	36.26	35.87	40.93
✓		✓	52.34	47.24	43.79	41.62	41.15	39.68	38.69	37.34	42.73
✓	✓		68.04	**63.94**	60.22	**58.00**	**55.44**	51.51	51.88	52.74	57.72
✓	✓	✓	**68.71**	63.87	**60.94**	57.98	55.27	**52.41**	**52.68**	**53.12**	**58.12**

Table 2. **Influence of OrCo loss components.** PSCL denotes perturbed supervised contrastive loss, CE denotes cross-entropy, ORTH denotes orthogonality loss. See figure 3 for visualization of each component. Ablation study on mini-ImageNet.

Sampling	aHM
Rand	57.23
Orth	**58.12**

Table 3. **Importance of explicit orthogonality loss for pseudo-target generation.** Rand denotes random sampling from normal distribution.

ble 1 presents the results obtained on the mini-ImageNet dataset, while figure 4 illustrates the evaluation results on the CUB200 and CIFAR100 datasets. Our method demonstrates superior performance across all three datasets, surpassing previous state-of-the-art methods by a significant margin, particularly achieving improvements of 9.4% and 5.5% on CIFAR100 and mini-ImageNet, respectively. Notably, the effectiveness of OrCo is consistently evident across all incremental sessions.

In addition to reporting results for the standard FSCIL methods, we also present the performance of the Imprinted Weights method (IW) [30] based on our model pretrained during Phase 1. The robust performance of this method indicates the efficacy of our pretraining strategy in facilitating effective transferability to downstream tasks, such as incremental few-shot learning sessions, thereby addressing the intransigence problem. We present a detailed breakdown of each session in section 2 of the supplement.

4.3. Analysis

To validate the effectiveness of each component of our framework, in this section, we show an analysis based on the mini-ImageNet dataset.

OrCo loss. We assess the efficacy of the components comprising our OrCo loss in table 2. The OrCo loss consists of three integral components illustrated in figure 3: the cross-entropy loss (CE), the orthogonality loss (ORTH), and the perturbed supervised contrastive loss (PSCL). We observe that CE struggles to generalize on underrepresented incremental classes. On the contrary, PSCL enhances the robust SCL approach with pseudo target perturbations and provides better class separation. PSCL, on its own, shows steady generalization, with only a 14.83% drop in harmonic mean. CE, however, while starting strong, ultimately becomes biased towards base classes, leading to a significant 49.26% drop in harmonic mean. By integrating the dynamic yet fundamentally discerning features of CE with the stability offered by PSCL, a significant enhancement in harmonic mean is achieved. Moreover, the orthogonality loss (ORTH) consistently improves performance (+0.4%). Its integration into our loss formulation further underscores the significance of orthogonality.

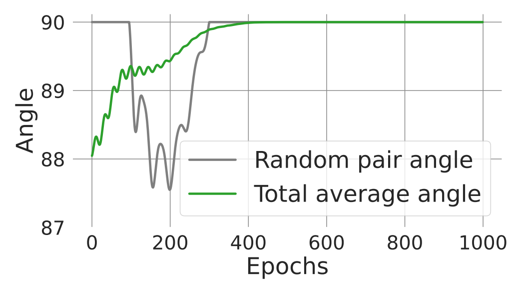

Figure 5. **Measurement of angle during orthogonality optimization.** The green curve corresponds to the evolution of the average angle between all pairs during the optimization. The gray curve shows measurements of random pairs at each epoch.

Influence of mutually orthogonal pseudo-targets. We note that independent randomly sampled vectors from a Gaussian distribution $\mathcal{N}(0, 1)$ are theoretically orthogonal on the surface of a unit sphere (refer to the supplement for a discussion). However, in practice, we observe only a near-orthogonal behavior, as illustrated in our training curve for orthogonal pseudo-target generation in figure 5. To examine the impact of a perfectly orthogonal target space, we assess our method using both randomly sampled targets from a Gaussian distribution and our generated orthogonal targets, as presented in table 3. We observe a 0.89% improvement in performance when explicit orthogonality constraints are applied. This finding suggests that an aligned and orthogonal feature space is more effective in addressing data imbalances between base and incremental sessions. Consequently, we incorporate orthogonality as a fundamental principle in our framework, recognizing its significant role in enhancing the overall effectiveness of our model.

Pseudo-targets perturbations. OrCo relies on perturba-

Perturbation	$FP_{inc} \downarrow$	$Sim_{cls} \downarrow$	$Sim_{cls \rightarrow target} \downarrow$	$HM_8 \uparrow$
w/o	66.5	0.105	0.013	20.14
$\mathcal{N}$	54.6	**0.002**	**0.006**	50.23
$\mathcal{U}$	**52.5**	0.011	**0.006**	**53.12**

Table 4. **Influence of perturbations in PSCL.** Comparison of our PSCL loss with or without perturbations of pseudo-targets. $\mathcal{N}$, $\mathcal{U}$ denotes Gaussian and Uniform distributions, respectively, from which λ, as in equation 4, sampled during training. FP_{inc} refers to the False Positive rate among all incremental classes. Sim_{cls} computes the average pairwise cosine similarity between all class pairs. $Sim_{cls \rightarrow target}$ indicates the pairwise cosine similarity between classes and unassigned target pairs over all sessions. HM_8 refers to the 8-th and final session harmonic mean.

tions of fixed pseudo-targets to introduce a margin between previously encountered and incoming classes. We compare OrCo against a variant where the contrastive loss does not receive any pseudo-targets' perturbations (w/o). In contrast to this, our perturbation schemes with sampling λ, as in equation 4, from Gaussian ($\mathcal{N}$) and uniform ($\mathcal{U}$) distributions consistently enhance the final session harmonic mean (HM_8) by over 30%, as shown in table 4.

For a detailed evaluation of our method, we employ false positive and cosine similarity analyses. By measuring the false positive rate within only incremental classes (FP_{inc}), we observe improved separation between the few-shot classes with the perturbed objective.

Subsequently, we calculate the average inter-class cosine similarity (Sim_{cls}) for all base and few-shot incremental classes, providing an indication of the spread of each class on the unit sphere. A lower value suggests more compact representations. Notably, we observe values at least 10 times lower for the training with perturbations. Lastly, we assess the availability of space around unassigned pseudo-targets ($Sim_{cls \rightarrow target}$) by computing the average similarity of all data features with respect to all unassigned targets. A higher average similarity corresponds to smaller margins between the features and the unassigned pseudo-targets. Table 4 illustrates that perturbations indeed increase the margin around unassigned pseudo-targets. Further discussions can be found in the supplement.

Influence of pretraining. To evaluate our pretraining strategy, we compare it against cross entropy (CE) and standard supervised contrastive loss (SCL) [17]. As shown in table 5, the addition of self-supervised contrastive loss (SCL+SSCL) to the pretraining session significantly enhances generalization on unseen data, showcasing improved transfer capabilities, which aligns with previous findings [5, 15]. Additionally, we present the accuracy on the validation set for the base classes D^0 immediately after the pretrain phase for each strategy.

Pretrain Strategy	Phase 1:Accuracy	aHM
CE	85.70	55.20
SCL	85.18	57.38
SCL + SSCL (Ours)	**85.95**	**58.12**

Table 5. **Influence of pretraining.** aHM denotes average harmonic mean. CE is cross-entropy, SCL is supervised countrastive loss, and SSCL is self-supervised contrastive loss.

Fine-tuned params	Performance Decay $\downarrow$	Base Decay $\downarrow$
f, g	28.94	20.52
g	**26.99**	**17.83**

Table 6. **Influence of frozen paramters.** Analysing of catastrophic forgetting when 1) fine-tuning the entire model (f, g) and 2) fine-tuning only the projection head (g).

While all strategies exhibit close to 85% accuracy on the base validation set, our approach yields a 0.74% higher average harmonic mean compared to *SCL* and a notable 2.92% improvement over *CE*. The significance of this lies in the fact that our frozen backbone network, maintained during incremental sessions, is capable of producing strong and unique features even for unseen classes.

Frozen parameters. Table 6 illustrates how OrCo effectively addresses catastrophic forgetting by adopting a strategy of freezing the backbone and training only the projection head. The observed overall performance decay, along with a 2.7% greater loss of base accuracy across all sessions, demonstrates favorable outcomes for decoupling the learning process after Phase 1.

5. Conclusion

This paper introduced OrCo method to boost the performance of FSCIL by addressing its inherent challenges: catastrophic forgetting, overfitting, and intransigence. The OrCo framework is a novel approach that tackles these issues by leveraging features' mutual orthogonality on the representation hypersphere and contrastive learning. By combining supervised and self-supervised contrastive learning during pretraining, the model captures diverse semantic information crucial for novel classes with limited data, implicitly addressing the intransigence challenge. Employing the proposed OrCo loss during subsequent incremental sessions ensures alignment with the generated fixed pseudo-targets, maximizing margins between classes and preserving space for incremental data. This comprehensive approach not only enhances feature space generalization but also mitigates overfitting and catastrophic forgetting, marking steps toward improving the practical value of incremental learning methods in real-world applications.

References

[1] Touqeer Ahmad, Akshay Raj Dhamija, Steve Cruz, Ryan Rabinowitz, Chunchun Li, Mohsen Jafarzadeh, and Terrance E. Boult. Few-shot class incremental learning leveraging self-supervised features. In *CVPR Workshops*, 2022. 2

[2] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *ICCV*, 2019. 1, 2

[3] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 2, 4

[4] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018. 1, 2

[5] Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *ICML*. PMLR, 2022. 2, 4, 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 2020. 2

[7] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 1

[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. 1, 2, 4

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR, 2017. 2

[10] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1

[11] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010. 2

[12] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *CVPR*, 2022. 1, 2, 6

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2

[14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 1, 2

[15] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *ICCV*, 2021. 2, 4, 8

[16] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. In *ICLR*, 2018. 1

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NIPS*, 2020. 2, 4, 8

[18] Do-Yeon Kim, Dong-Jun Han, Jun Seo, and Jaekyun Moon. Warping the space: Weight space rotation for class-incremental few-shot learning. In *ICLR*, 2022. 1, 2

[19] 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*, 2017. 1, 2

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 4

[22] A Kukleva, H Kuehne, and B Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. in 2021 ieee. In *ICCV*, 2021. 1, 2, 4, 6

[23] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, pages 6918–6928, 2022. 3

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 2017. 1, 2

[25] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018. 1, 2

[26] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *AAAI*, 2021. 1, 2

[27] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Elsevier, 1989. 1

[28] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[30] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *CVPR*, 2018. 2, 6, 7

[31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 1, 2

[32] 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. *IJCV*, 115, 2015. 6

[33] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable

framework for continual learning. In *ICML*. PMLR, 2018. 1, 2

[34] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NIPS*, 30, 2017. 1, 2

[35] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *CVPR*, 2020. 1, 2, 6

[36] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 2024. 2

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NIPS*, 29, 2016. 2

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[39] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019. 1, 2, 4

[40] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *ICLR*, 2023. 1, 2, 6

[41] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 6

[42] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *CVPR*, 2021. 1, 2, 6

[43] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, 2020. 1, 2

[44] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE TPAMI*, 2021. 1, 2

[45] Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *CVPR*, 2023. 2, 6

[46] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *CVPR*, 2022. 1, 2, 6

[47] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE TPAMI*, 2022. 6

[48] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, 2021. 1, 2