DRESS: Disentangled Representation-based Self-Supervised Meta-Learning for Diverse Tasks

Wei Cui
Layer 6 AI
wei@layer6.ai

Yi Sui
Layer 6 AI
amy@layer6.ai

Jesse C. Cresswell
Layer 6 AI
jesse@layer6.ai

Keyvan Golestan
Layer 6 AI
keyvan@layer6.ai

Abstract

Meta-learning represents a strong class of approaches for solving few-shot learning tasks. Nonetheless, recent research suggests that simply pre-training a generic encoder can potentially surpass meta-learning algorithms. In this paper, we first discuss the reasons why meta-learning fails to stand out in these few-shot learning experiments, and hypothesize that it is due to the few-shot learning tasks lacking diversity. Furthermore, we propose *DRESS*, a task-agnostic Disentangled REpresentation-based Self-Supervised meta-learning approach that enables fast model adaptation on highly diversified few-shot learning tasks. Specifically, DRESS utilizes disentangled representation learning to create self-supervised tasks that can fuel the meta-training process. We validate the effectiveness of DRESS through experiments on few-shot classification tasks on datasets with multiple factors of variation. Through this paper, we advocate for a re-examination of proper setups for task adaptation studies, and aim to reignite interest in the potential of meta-learning for solving few-shot learning tasks via disentangled representations.

1 Introduction & Background

Few-shot learning [1] emphasizes the ability to quickly learn and adapt to new tasks, and is regarded as one of the trademarks of human intelligence. In the pursuit of few-shot learning, meta-learning approaches have been widely explored [2–4], as they enable models to *learn-to-learn*. Nonetheless, recent research [5] suggests that a generic encoder is sufficient to support highly competitive performance on few-shot learning tasks. One can simply train an encoder on the unified set aggregating samples over meta-training tasks. A linear layer is then added on top of the encoder and is fine-tuned using the few-shot support samples to adapt to new tasks. Despite completely ignoring information about task identity, this simple scheme can achieve better results than meta-learning methods [5]. This finding is surprising, perhaps even confusing, since it suggests that the identities of and distinctions between individual training tasks are irrelevant to achieving high learning performance.

We believe this observation can be explained by the fact that diversity among tasks is lacking in many few-shot learning benchmarks. Particularly, for popular few-shot learning datasets such as Omniglot [6], *mini*ImageNet [7], and CIFAR-FS [8], the tasks differ only by their targets being from disjoint sets of object classes. These few-shot learning tasks are essentially all object classification tasks. There is one strategy for solving each of these tasks simultaneously that works without the need to individually identify each task: simply compare the main object in each query image to the objects in the few-shot support images, and assign the class label according to the maximally similar support image. Contrastive learning coupled with semantic-preserving augmentations can directly implement this strategy of aligning the main objects while discarding other factors (such as orientation and background) [9]. Therefore, on these specific benchmarks, it is no wonder that a single encoder aided by contrastive learning can compete against meta-learning methods while ignoring the essential information of task identities and distinctions.

The notion of task diversity has been studied previously [10–12]. Recently, [13] conducted thorough experiments suggesting that existing meta-learning methods show very slight improvements over the

NeruIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice.

pre-training and fine-tuning scheme on tasks with higher Task2Vec diversity coefficients [11, 12]. Nonetheless, to the best of our knowledge, the intuition behind task diversities and the performance of few-shot learning has yet to be discussed, nor has any meta-learning approach explicitly exploited the idea of diversifying meta-training tasks for boosting the fast adaptation ability of a model.

In this paper, we focus on fast adaption to few-shot learning tasks with diversified and distinctive natures. We consider tasks beyond main object classification, namely identifying orientation, background angle or color, attributes of secondary objects in the image, and so on. More importantly, to fully examine few-shot learning ability, the model should be *agnostic* to the nature of the evaluation tasks. Such setups can reveal the model's true capability to learn strictly from the few-shot samples, with *task identification* being an essential part of learning.

For effective meta-learning under high task diversities, we bridge the idea of disentangled representation learning with self-supervised meta-learning, and propose our approach *DRESS — task-agnostic Disentangled REpresentation-based Self-Supervised meta-learning*. Specifically, we utilize an encoder trained to compute disentangled representations for obtaining encodings of inputs. We then perform clustering independently within each disentangled latent dimension, and use the cluster identities to define pseudo-classes. Lastly, we construct a set of self-supervised few-shot classification tasks based on these pseudo-classes for each latent dimension. With the disentangled latent dimensions representing distinct factors and attributes within the input images, the constructed few-shot learning tasks are highly diversified. Using these tasks for meta-training, the model can learn to quickly adapt to various unknown tasks, regardless of which aspects of inputs the tasks focus on. We conduct experiments on image datasets containing multiple varying factors, beyond the class of the main object. Our results suggest that DRESS enables few-shot learning performance that can surpass existing methods and approaches the upper-bound of supervised baselines.

2 Self-Supervised Meta-Learning with Disentangled Representations

Self-Supervised Meta-Learning Researchers have recently been interested in the possibilities of unsupervised or self-supervised meta-learning [14–19]. As an example, [14] proposes an unsupervised task construction approach using an encoding-then-clustering scheme. While promising results are obtained, [14] does not explicitly address the problem of adaptation under high task diversities. Our motivation is to generalize to diverse tasks without the costly target collection process required in these previous approaches. Under high task diversity, labels in the meta-training set might *mislead* the model to exclusively capture information irrelevant or unsuitable for solving the meta-testing tasks. With the general assumption that the meta-testing tasks should be *agnostic* to the model, a self-supervised meta-training process is preferable, as the model will not fixate on the specific labels of a supervisedmeta-training process.

Disentangled Representation Learning Disentangled representation learning has been investigated in the context of generative modeling [20–24]. One of the main metrics to judge disentangled representations by is *completeness* [25]. High completeness means the representation captures more of the variation within the input distribution. For complex images, factors of variations include the main object, and in addition, the background, the view angle, and so on. In this paper, we propose to bridge disentangled representation learning with self-supervised meta-learning. Specifically, we recognize the rich and diversified information disentangled representations provide, and utilize them to fuel the meta-training process.

DRESS: Disentangled Representation-based Self-Supervised Meta-Learning We now propose DRESS, our task agnostic Disentangled REpresentation-based Self-Supervised meta-learning approach. First, we collect all datapoints available for meta-training, and compute disentangled latent representations using an encoder, for example a factorized diffusion autoencoder (FDAE) [24]. Then, we cluster the datapoints along each disentangled latent dimension. Lastly, we construct self-supervised learning tasks using the cluster identities as the pseudo-class labels. We construct a large number of few-shot classification tasks under each disentangled latent dimension by first sampling a subset of cluster identities as classes, and then sampling a subset of datapoints under each class as the few-shot support samples and query samples. We provide the flow diagram illustration for DRESS in Figure 1. We also include the entire meta-learning pipeline illustration in Appendix F. As different dimensions within the disentangled representation depict distinct aspects of the input data, the sets of self-supervised tasks constructed from disentangled dimensions are naturally diversified, requiring distinct decision rules to solve. When using these tasks for meta-training, the model can

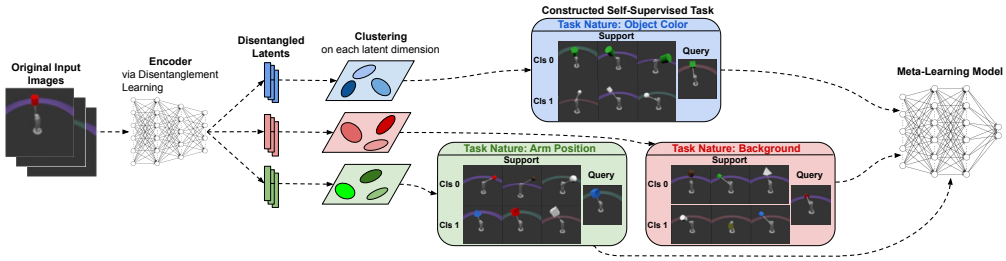

Figure 1: DRESS creates diversified self-supervised meta-training tasks through disentanglement learning. The images are first encoded into disentangled latent representations. Sets of clusters are formed on each latent dimension individually. Each set of clusters form pseudo classes for a distinct set of self-supervised classification tasks with their own unique nature, under each latent dimension.

digest each factor of variation within the data, and therefore learns to adapt to unseen few-shot tasks regardless of their contexts, natures, and meanings.

3 Experiments & Analysis

Datasets and Task Setups To study highly diversified tasks, we utilize two image datasets with multiple independently varying factors, Shapes3D [26] and MPI3D [27], which are often used in research on generative modeling and disentangled representations. We provide the details of all varying factors for Shapes3D and MPI3D in Appendix A. For meta-evaluation as well as meta-training under supervised baselines (as introduced below), we prepare supervised few-shot learning tasks, with the detailed descriptions also provided in Appendix A. These few-shot learning setups reflect the realistic scenario where the tasks that the model needs to adapt to are agnostic to the model during its training process. We emphasize that we choose not to experiment on popular datasets including Omniglot, *mini*ImageNet, and CIFAR-FS, due to their lack of task diversities among meta-training and meta-testing stages, as we have elaborated in section 1.

Implementation Details of DRESS We use the FDAE architecture [24] as the encoder for obtaining disentangled representations, which is trained from scratch on meta-training images from Shapes3D or MPI3D. The FDAE encoder computes a pair of codes for each visual concept, the content code and the mask code. We regard these codes as two independent spaces, each with dimensionality given by the FDAE encoding vector. Hence, the 6 visual concepts in Shapes3D images and 7 visual concepts in MPI3D images lead to 12 and 14 disentangled latent dimensions for each dataset respectively. Under each latent dimension (i.e. a vector for each image), we first apply PCA to reduce the dimensionality, and then compute 200 clusters via K-Means, which serve as the pseudo-classes for self-supervised meta-training tasks. Hence, each disentangled latent dimension corresponds to a distinct set of few-shot classification tasks with their own unique nature.

Baseline Methods We compare DRESS to several popular baseline methods on few-shot learning.

Pre-training and Fine-tuning We implement the encoder-based pre-training and fine-tuning method as described in [5], using SimCLR [28] with its standard image augmentation pipeline for pre-training. The details are elaborated on in Appendix D.

Supervised Meta-Learning We evaluate three supervised meta-learning baselines, with increasing knowledge of ground-truth factors and setups for tasks: *Supervised-Original*, *Supervised-All*, and *Supervised-Oracle*. Exact definitions of these baselines are provided in Appendix B. *Supervised-Original* is a commonly used baseline for meta-learning [14], while we have designed the two more powerful baselines to reflect the upper bounds of what may be achieved with meta-learning in principle. Specifically, *Supervised-All* serves as the upper bound on performance attainable when the evaluation tasks are agnostic to the model, while *Supervised-Oracle* serves as the ultimate performance upper bound when both the ground-truth factors as well as the natures of the evaluation tasks (i.e. which factors the evaluation tasks focus on) are perfectly known.

Unsupervised & Self-Supervised Meta-Learning We utilize *CACTUS* [14] as an unsupervised meta-learning baseline, implemented with two encoders: DeepCluster [29] trained from scratch on Shapes3D or MPI3D; and off-the-shelf DINOv2 [30]. We refer to these two variants as *CACTUS-DeepCluster* and *CACTUS-DINOv2*. Detailed settings for them are provided in Appendix C. While there are many more methods which could serve as relevant baselines, e.g. [16, 17, 19], their use of

Table 1: Few-shot learning classification accuracies, aggregated over 4 trials with different seeds, with each trial conducted over 1000 meta-testing few-shot learning tasks.

Method	Shapes3D	MPI3D-Easy	MPI3D-Hard
Supervised-Original	$62.03\% \pm 1.55\%$	$57.75\% \pm 0.46\%$	$63.27\% \pm 1.25\%$
Supervised-All	$99.93\% \pm 0.02\%$	$99.29\% \pm 0.29\%$	$91.03\% \pm 1.70\%$
Supervised-Oracle	$99.97\% \pm 0.02\%$	$100.00\% \pm 0.00\%$	$99.42\% \pm 0.11\%$
Few-Shot Direct Adaptation	$65.70\% \pm 2.05\%$	$60.59\% \pm 0.29\%$	$62.27\% \pm 0.28\%$
Pre-Training and Fine-Tuning	$57.88\% \pm 2.19\%$	$92.93\% \pm 0.48\%$	$79.50\% \pm 0.76\%$
CACTUS-DeepCluster	$86.81\% \pm 0.68\%$	$84.95\% \pm 0.56\%$	$72.77\% \pm 0.97\%$
CACTUS-DINOv2	$80.62\% \pm 0.25\%$	$94.39\% \pm 0.44\%$	$81.92\% \pm 0.39\%$
DRESS	$93.05\% \pm 0.18\%$	$99.94\% \pm 0.03\%$	$84.95\% \pm 0.50\%$

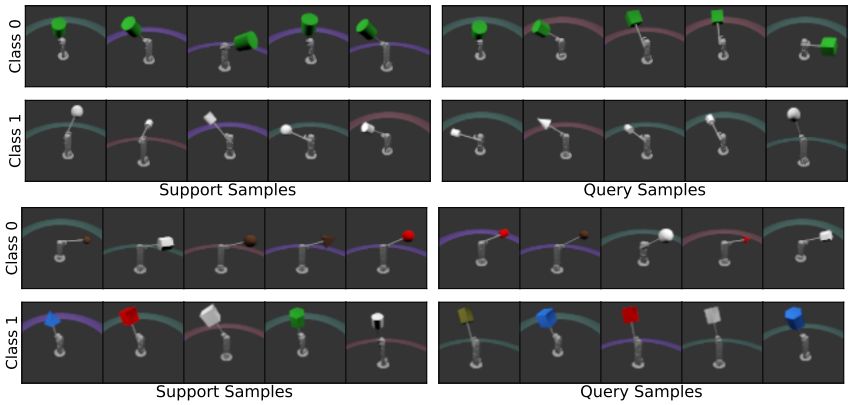

Figure 2: Two self-supervised tasks constructed by DRESS on MPI3D. The top task focuses on the object color; while the bottom task focuses on the robot arm angle. In particular, in the bottom task, DRESS is able to incorporate the effect of camera height when comparing the robot arm angles.

data augmentation and pre-training techniques (i.e. contrastive learning) makes them close in nature to our Pre-training and Fine-tuning baseline as well as CACTUS-DINOv2, since DINOv2 is already a strong self-supervised image encoder. Lastly, we use *Few-Shot Direct Adaptation* as a lower bound of what should be achievable with the model directly optimized on the few-shot support samples from each meta-evaluation task.

Results and Analysis We present few-shot classification accuracies in Table 1. On these datasets, DRESS consistently achieves the best few-shot adaptation performance (except for Supervised-All and Supervised-Oracle as upper bounds). The performance of Supervised-Original is unimpressive, indicating that meta-training targets could mislead a supervised model when adapting to highly diversified tasks as we discussed in section 2. In contrast to [5], Pre-training and Fine-Tuning is not on par with meta-learning approaches, due to the more challenging and diverse tasks we benchmark on. CACTUS shows varying results across datasets with different encoders, reflecting the importance of the latent representations in task construction. As DRESS uses disentangled representation learning to construct diversified pre-training tasks, it obtains superior results across these datasets and task setups. We provide visualizations of two tasks constructed by DRESS in Figure 2 which require the model to learn to identify the object color, and the robot arm angular position respectively. In Appendix E, we visualize more tasks constructed by DRESS focusing on other factors within MPI3D.

4 Conclusion

We surfaced an issue in popular few-shot learning benchmarks: the tasks are not diversified enough to truly test model adaptation ability. Instead, few-shot learning tasks that cover various aspects of the inputs and require distinct learning rules serve as more informative benchmarks. We proposed a self-supervised meta-learning approach that harnesses the expressiveness of disentangled representations to construct self-supervised tasks. Our approach enables models to acquire broad knowledge on underlying factors in the dataset, and quickly adapt to tasks of various natures.

References

[1] Yaqing Wang, Quanming Yao, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[3] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[4] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *The International Conference on Learning Representations (ICLR)*, 2017.

[5] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020.

[6] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In *Conference of the Cognitive Science Society (CogSci)*, 2011.

[7] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[8] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR)*, 2019.

[9] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv:2304.12210*, 2023.

[10] Amir R. Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[12] Brando Miranda, Patrick Yu, Yu-Xiong Wang, and Sanmi Koyejo. The curse of low task diversity: On the failure of transfer learning to outperform maml and their empirical equivalence. *arXiv preprint arXiv:2208.01545*, 2022.

[13] Brando Miranda, Patrick Yu, Saumya Goyal, Yu-Xiong Wang, and Sanmi Koyejo. Is pre-training truly better than meta-learning? *arXiv preprint arXiv:2306.13841*, 2023.

[14] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *The International Conference on Learning Representations (ICLR)*, 2019.

[15] Antreas Antoniou and Amos Storkey. Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. *arXiv preprint arXiv:1902.09884*, 2019.

[16] Siavash Khodadadeh, Ladislau Bölöni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[17] Siavash Khodadadeh, Sharare Zehtabian, Saeed Vahidian, Weijia Wang, Bill Lin, and Ladislau Boloni. Unsupervised meta-learning through latent-space interpolation in generative models. In *The International Conference on Learning Representations (ICLR)*, 2021.

[18] Hui Xu, Jiaxing Wang, Hao Li, Deqiang Ouyang, and Jie Shao. Unsupervised meta-learning for few-shot learning. *Pattern Recognition*, 2021.

[19] Huiwon Jang, Hankook Lee, and Jinwoo Shin. Unsupervised meta-learning via few-shot pseudo-supervised contrastive learning. In *The International Conference on Learning Representations (ICLR)*, 2023.

[20] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *The International Conference on Learning Representations (ICLR)*, 2017.

[21] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.

[22] Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. DisDiff: Unsupervised disentanglement of diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[23] Kyle Hsu, Jubayer Ibn Hamid, Kaylee Burns, Chelsea Finn, and Jiajun Wu. Tripod: Three complementary inductive biases for disentangled representation learning. In *International Conference on Machine Learning (ICML)*, 2024.

[24] Ancong Wu and Wei-Shi Zheng. Factorized diffusion autoencoder for unsupervised disentangled representation learning. In *The Association for the Advancement of Artificial Intelligence Conference (AAAI)*, 2024.

[25] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *The International Conference on Learning Representations (ICLR)*, 2018.

[26] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[27] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[29] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

A Dataset Descriptions

A.1 Shapes3D

Shapes3D contains 480,000 images, of which we use 400,000 images for meta-training and 50,000 images for meta-testing, leaving 30,000 images for meta-validation. Each image has a resolution of 64×64 pixels with RGB color channels.

The images in the dataset incorporate 6 factors of variations, as detailed in Table 2.

Table 2: Factors of Variation in Shapes3D

Attribute Name	Number of Values	Constructed Tasks
Floor Hue	10	Meta-Test
Wall Hue	10	Meta-Test
Object Hue	10	Meta-Train
Scale	8	Meta-Train
Shape	4	Meta-Train
Orientation	15	Meta-Test

A.2 MPI3D

MPI3D consists of four dataset variants. We utilize the *MPI3D_toy* dataset containing simplistic rendered images. Throughout the paper, we refer to this dataset simply as MPI3D. The dataset contains 1,036,800 images, of which we use 1,000,000 images for meta-training and 30,000 images for meta-testing, leaving 6,800 images for meta-validation. Each image has a resolution of 64×64 pixels with RGB color channels.

The images in the dataset incorporate 7 factors of variations, as detailed in Table 3. We note that for the two factors *horizontal axis* and *vertical axis*, denoting the robot arm's angular position, the ground truth labels for each are based on a 40-interval partition of the entire 180-degree anglular range, leading to a mere 4.5-degree maximum angle difference for two different factor values. In our experiments, we re-group the partitions into 10 intervals for each of the two axes, leading to a 18-degree maximum angle difference between two factor values.

Table 3: Factors of Variation in MPI3D

Attribute Name	Number of Values	Constructed Tasks (Easy)	Constructed Tasks (Hard)
Object Color	6	Meta-Train	Meta-Train
Object Shape	6	Meta-Train	Meta-Train
Object Size	2	Meta-Train	Meta-Train
Camera Height	3	Not Used	Meta-Test
Background Color	3	Not Used	Meta-Test
Horizontal Axis	40	Meta-Test	Not Used
Vertical Axis	40	Meta-Test	Not Used

A.3 Attribute Splits for Supervised Few-Shot Learning Tasks

We prepare few-shot binary classification tasks, specifically, *two-way five-shot tasks*, based on the ground-truth factor values, following the procedure in [14]: select a subset of attributes and two value combinations of these attributes, then assign the positive class (and negative class) to images whose attributes match the first (and second) value combination. We also prepare five query samples under each class (i.e. a total of ten query samples to be classified by the model). Note that we do not provide the model with the class-distribution information over these query samples.

For the attribute subset selection, we partition the entire set of factors into disjoint meta-training sets and meta-testing sets. The detailed split for both Shapes3D and MPI3D on each attribute is listed above in Table 2 and Table 3 respectively.

Specifically for MPI3D, we conduct experiments on two setups: *MPI3D-Easy* and *MPI3D-Hard*. For MPI3D-Easy, the meta-testing tasks focus on identifying the background and camera height attributes, while for MPI3D-Hard, the meta-testing tasks focus on predicting the horizontal and vertical robot arm angular positions (which are also perturbed visually at various camera heights).

B Supervised Meta-Learning Baselines

In this section, we provide the complete description of the three supervised meta-learning baselines adopted in our experiments.

- *Supervised-Original*: Only use the ground-truth factors allocated for meta-training to prepare few-shot learning tasks to meta-train the model.
- *Supervised-All*: Use all the ground-truth factors to prepare few-shot learning tasks to meta-train the model.
- *Supervised-Oracle*: Use the ground-truth factors allocated for meta-testing to prepare few-shot learning tasks to meta-train the model.

These method progressively increase the amount or relevance of information available to the model. Supervised-Original must learn to generalize from a limited set of factors to completely new factors at meta-test-time. Supervised-All has strictly more information than the other two, but must be prepared to adapt to any of them meta-test-time. Supervised-Oracle is given exactly the factors that it will be tested on, so generalization to new factors is not required for good performance. Hence, it is feasible for either Supervised-All or Supervised-Oracle to come out ahead as the upper-bound on performance. For instance, if there are similar or overlapping factors across meta-train and meta-test, Supervised-All may benefit for seeing all of it.

C Setups for All Meta-Learning Methods

For DRESS as well as each meta-learning baselines, we generate self-supervised few-shot learning tasks, following the same format as the supervised tasks specified in Appendix B: two-way five-shot, with five query samples per class.

For each meta-learning method (including the supervised meta-learning methods), we use MAML [2] as the meta-optimization engine, with a convolutional neural network of identical specification as the base learner, for fair comparisons between the methods. For each method, we use the setup for meta-training and meta-evaluation as listed in Table 4.

Table 4: Few-Shot Learning Setup

Setting	Value
Tasks per Meta-Training Epoch	8
Meta-Training Epochs	30,000
Tasks in Meta-Evaluation	1,000
Gradient Descent Steps in Adaptation	5
Adaptation Step Learning Rate	0.05
Meta-Optimization Step Learning Rate	0.001

We summarize in Table 5 all hyper-parameter values of meta-learning baselines. We note that for the DeepCluster encoder, PCA is applied on its output to reach the number of latent dimensions as listed.

Table 5: Few-Shot Learning Setup

Setting	CACTUS-DeepCluster	CACTUS-DINOv2
Latent Dimensions	256	384
Randomly Scaled Latent Spaces	50	50
Clusters Over Each Latent Space	300	300

D Setups for Pre-training and Fine-tuning

For pre-training, we use an encoder backbone that shares the same architecture as the ResNet-18 [31] backbone used for FDAE. After the pre-training, a trainable linear layer is attached on top of the encoder for the adaptation process on evaluation tasks. The encoder is frozen throughout the adaptation process. We include the details for this approach in Table 6. Note that we do not use a supervised loss in pre-training in order to avoid the encoder focusing only on tasks that are irrelevant to the meta-evaluation tasks, as we have discussed in section 2.

Table 6: Pre-Training and Fine-Tuning Setup

Setting	Value
Pre-Training Epochs	10
Tasks in Meta-Evaluation	1000
Gradient Descent Steps in Adaptation	5

Regarding the number of epochs for pre-training, in the pre-training procedure, the entire set of meta-training image inputs are fed to the encoder (i.e. 400,000 images for Shapes3D; and 1,000,000 images for MPI3D). Therefore, with 10 epochs over the entire meta-training dataset, the number of forward-backward computations for optimizing the encoder already surpasses the models trained with the meta-learning methods.

E Additional Task Visualizations from DRESS

We provide more visualizations on self-supervised few-shot learning tasks generated by DRESS in Figure 3. These visualizations illustrate that the generated tasks cover multiple aspects and factors of variations within the dataset.

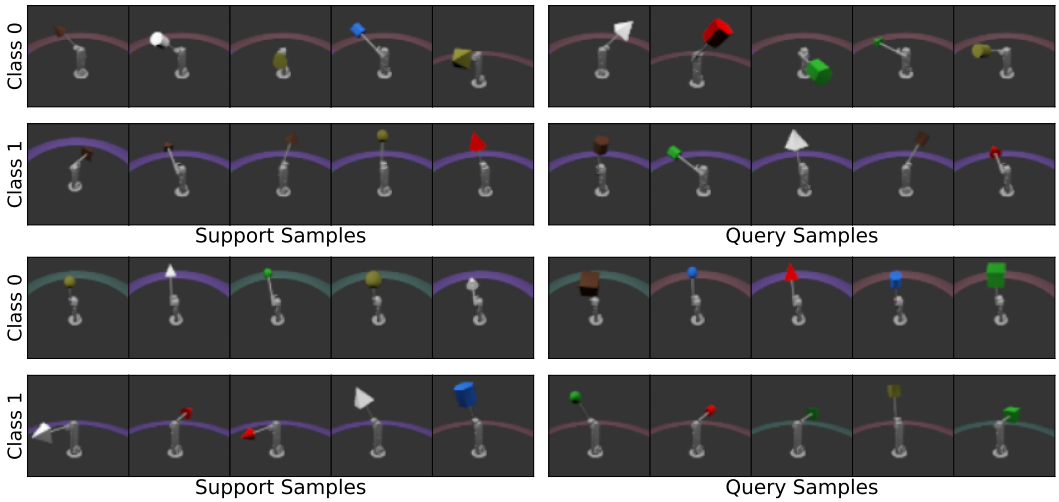

Figure 3: More self-supervised tasks constructed by DRESS. The top task focuses on the background color; while the bottom task focuses on the camera height.

F Meta-Learning on Few-Shot Learning Pipeline

We provide the visualization for the general pipeline on applying meta-learning to solve few-shot learning tasks in Figure 4.

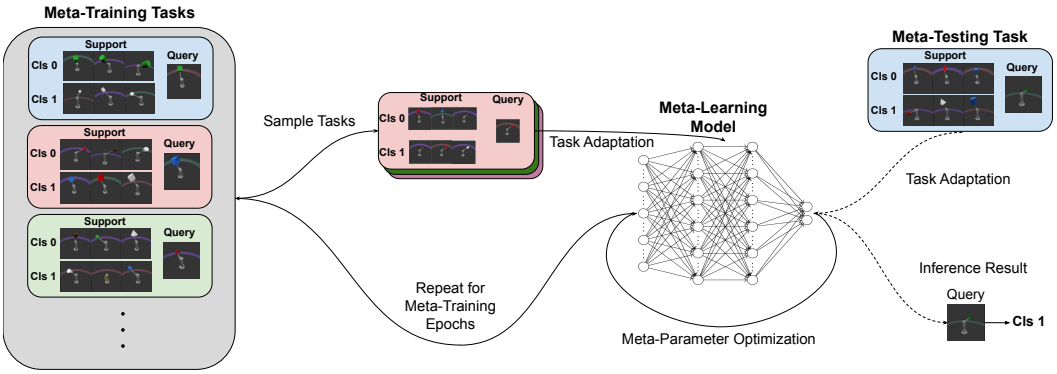

Figure 4: During meta-training stage, the model adapts on batches of sampled (self-supervised) tasks. The model's performance is optimized for meta-parameter optimization. After meta-training, the model can be quickly adapted to meta-testing tasks and perform few-shot inference.