ReFine: Re-randomization before Fine-tuning for Cross-domain Few-shot Learning

Jaehoon Oh∗
KAIST DS
Daejeon, Republic of Korea
jhoon.oh@kaist.ac.kr

Sungnyun Kim∗
KAIST AI
Seoul, Republic of Korea
ksn4397@kaist.ac.kr

Namgyu Ho∗
KAIST AI
Seoul, Republic of Korea
itsnamgyu@kaist.ac.kr

Jin-Hwa Kim
NAVER AI Lab
Sungnam, Republic of Korea
jlhwa.kim@navercorp.com

Hwanjun Song†
NAVER AI Lab
Sungnam, Republic of Korea
hwanjun.song@navercorp.com

Se-Young Yun†
KAIST AI
Seoul, Republic of Korea
yunseyoung@kaist.ac.kr

ABSTRACT
Cross-domain few-shot learning (CD-FSL), where there are few target samples under extreme differences between source and target domains, has recently attracted huge attention. Recent studies on CD-FSL generally focus on transfer learning based approaches, where a neural network is pre-trained on popular labeled source domain datasets and then transferred to target domain data. Although the labeled datasets may provide suitable initial parameters for the target data, the domain difference between the source and target might hinder fine-tuning on the target domain. This paper proposes a simple yet powerful method that re-randomizes the parameters fitted on the source domain before adapting to the target data. The re-randomization resets source-specific parameters of the source pre-trained model and thus facilitates fine-tuning on the target domain, improving few-shot performance.

CCS CONCEPTS
• Computing methodologies → Machine learning.

KEYWORDS
cross-domain, few-shot, transfer learning, re-randomization

1 INTRODUCTION
Few-shot learning (FSL) has become an attractive field of deep learning research to tackle problems with a small number of training samples [38]. In this setting, a model is typically pre-trained on a large source dataset comprised of base classes from the source domain and then transferred into the target dataset comprised of few samples from unseen novel classes. Studies on FSL have typically assumed that the base and novel classes share the same domain, and these have followed two research directions: meta-learning [9, 23, 25, 31] and fine-tuning [4, 7, 34].

However, the source dataset and the target dataset come from considerably different domains in many real-world scenarios [13, 24]. To tackle this problem, cross-domain few-shot learning (CD-FSL) has recently gained significant attention, exemplified by the introduction of the BSCD-FSL benchmark dataset [13]. This benchmark considers large-scale natural image datasets as source data and four different target datasets for evaluation, each with varying levels of similarity to the source data domain. It is shown that transfer learning approaches, where a pre-trained model on the source domain is fine-tuned on the target domain, overwhelm meta-learning approaches on BSCD-FSL [13].
In this regard, recent works have attempted to extract better representations during the pre-training phase by exploiting unlabeled data from the target domain [16, 22, 24] or reconstructing the images with an autoencoder to enhance the generalization of a model [18]. While these works focus on developing better pre-training methods, we suppose the fine-tuning phase is also a crucial research direction. Das et al. [6] were aware of the importance of fine-tuning for CD-FSL, however, their framework using a mask generator is highly complicated to use.

In this paper, we present a new perspective to tackle the domain gap issue in CD-FSL: not all the pre-trained parameters from the source domain are desirable on the target domain. We posit that parameters in deeper layers of a pre-trained feature extractor may be detrimental for target domain adaptation, as they contain domain-specific information belonging to the source domain. This is demonstrated in Figure 1, where we use fixed image features from different stages of a pre-trained backbone and analyze the change in few-shot performance. We observe different trends for same-domain and cross-domain scenarios. While accuracy increases consistently with feature depth in the same-domain case (the blue lines), the accuracy decreases when using features from the last stage in the cross-domain case (the red lines).

Motivated by these findings, we propose a novel method, ReFine (Re-randomization before Fine-tuning), where we re-randomize the top layers of the feature extractor after supervised training on the source domain, before fine-tuning on the target domain. This is effective for CD-FSL because it helps reduce the learning bias towards the source domain by simply re-randomizing the domain-specific layer. It can also be implemented by adding a few lines of code and can be easily combined with other recent CD-FSL methods. This simplicity and flexibility allows it to be easily adapted in practical uses for CD-FSL. Contrary to the prior works that have focused on improving universal representations during the pre-training phase [22, 24], our method focuses on removing source-specific features obtained during pre-training to aid the fine-tuning.

Our contributions are summarized as follows:

• We propose a simple yet effective algorithm called ReFine, which re-randomizes the fitted parameters on the source domain and then fine-tunes the partially re-randomized model. This puts forward a new perspective for adapting to novel classes of the target domain in CD-FSL.

• We demonstrate improved performance for CD-FSL when our re-randomization technique is used, and provide an in-depth analysis on where and how to re-randomize.

## 2 RELATED WORKS

**Few-shot learning (FSL)** has been studied in two research directions, meta-learning and fine-tuning. Regarding the meta-learning approach, a meta-trained model is evaluated after fast adaptation on a few train sets. The meta-training procedure resembles the episodic evaluating procedure. Meta-learning approaches include learning good initialized parameters [9, 10, 23, 34], a metric space [3, 31, 32, 36], and update rule or optimization [2, 11, 27]. Regarding the fine-tuning approach [4, 7, 34], a pre-trained model is typically evaluated after fine-tuning. During the pre-training procedure, the model is trained in a mini-batch manner.

---

**Cross-domain few-shot learning (CD-FSL)** addresses a problem when the source and target domains are extremely different, which is a more real-world scenario for FSL [13, 35]. Initially, Tseng et al. [35] proposed feature-wise transformation (FWT) that learns scale-and-shift meta-parameters using pseudo-unseen target data during meta-training. However, it showed poor performance on the recently released BSCD-FSL benchmark [13], consisting of four target datasets collected from different domains. In general, fine-tuning based approaches have been shown to outperform meta-learning based approaches such as FWT [13]. Therefore, recent CD-FSL studies have proposed their algorithms under a pre-training and fine-tuning scheme. These works have mainly focused on improving the pre-training phase, so that the pre-trained model is more suitable for adaptation to the target domain.

**Re-randomization** has been widely studied in the field of language tasks [33, 41], in particular related to BERT, which is one of the most popular fine-tuning based language models. An interesting observation from Zhang et al. [41] is that re-randomizing the topmost block in BERT increases the performance for downstream tasks by reducing the fine-tuning workload. Concurrently to this observation, Tamkin et al. [33] examined the relations between the partial re-randomization of BERT and transferability of the layers. Meanwhile, in a visual task, Alabdulmohsin et al. [1] showed that placing more emphasis on the early layers of a convolutional neural network helps improve generalization. There have been similar attempts in meta-learning based FSL, e.g., zeroing the context vector for adaptation in each new task [42] and setting the classifier weight to have the same row vector (for any-shot problem) [8]. However, to the best of our knowledge, our work is the first to investigate the impact of re-randomization in fine-tuning based approaches for better CD-FSL.

Although some literature use the term re-initialization, we distinguish it from re-randomization because re-initialization reverts the values to the previously initialized ones. Refer to [40] for a formal definition. For a more concrete comparison, we have also dealt with re-initialization in Section 4.4.
3 REFINE: RE-RANDOMIZATION BEFORE FINE-TUNING

The objective of fine-tuning based CD-FSL algorithms is to learn a backbone \( f \) on the source data \( D_B \) with base classes \( C_B \), extracting meaningful representations on the target data \( D_N \) with novel classes \( C_N \), where \( C_B \cap C_N = \emptyset \). However, because there is no access to target data, the pre-trained model is biased towards the source domain, especially in the upper layers that pertain to classification of base classes. To mitigate this, we re-randomize the upper layers of the pre-trained backbone \( f \) to reset source-fitted parameters, depicted in Figure 2. Specifically, the weights of convolutional layers are re-randomized to uniform distributions [14]. The scaling and shifting parameters of batch normalization layers are reset to ones and zeros, respectively.

The reason why upper layers of the backbone \( f \) are re-randomized is that more domain-specific representations are extracted as the depth increases in convolutional neural networks [1, 19, 26, 39]. Re-randomization of upper layers helps the training loss to escape from local minima attributed to \( D_B \) and allows bottom-level layers to be sufficiently updated, alleviating the gradient vanishing problem [17, 29]. This is in line with previous works which show that representation change is beneficial for CD-FSL [23, 35].

Finally, fine-tuning and evaluation are performed with episodes, each representing distinct tasks, sampled from the labeled target data \( D_N \). Each episode consists of a support set \( D_s \), used to fine-tune the partially re-randomized pre-trained model, and a query set \( D_q \), used to evaluate after the fine-tuning. To sample an episode \((D_s, D_q)\), \( n \) classes are first selected from \( C_N \), and subsequently, \( k \) and \( k_q \) samples are selected per class for support and query sets, respectively, where \( n = 5 \) and \( k \in \{1, 5\} \) in general.

4 EXPERIMENTS

We introduce the experimental setup in Section 4.1 and compare ReFine (ours) with two baselines in Section 4.2: (1) Linear is a linear probing method to fine-tune only the classifier layer; (2) Transfer is a transfer learning method to fine-tune the entire network without using re-randomization\(^2\). We further investigate where and how to re-randomize in Section 4.3 and Section 4.4, respectively.

4.1 Experimental Setup

Datasets. For the source domain dataset, we use miniImageNet (miniIN) [36] and tieredImageNet (tieredIN) [28]. For the target domain, we use the BSCD-FSL benchmark [13], which consists of four different datasets: CropDisease [21], EuroSAT [15], ISIC [5], and ChestX [37], in order of similarity to miniIN.

Backbone and Training Setup. We use ResNet10 for miniIN and ResNet18 for tieredIN. Figure 2 describes the ResNet10 backbone. A family of ResNet consists of one stem module and four stages. The stem module consists of Conv-BN-ReLU-MaxPool layers. Each stage includes one or more convolution blocks, where resolution is halved and the number of channels is doubled in the first block, and they are maintained in the following blocks. For the pre-training and fine-tuning setups, we follow Guo et al. [13].

\(^2\)Many meta-learning based approaches such as MAML, ProtoNet, ProtoNet+FWT, and MetaOptNet have worse performance than Transfer, which is shown in [13].

4.2 Performance Comparison

Table 1 describes the 5-way \( k \)-shot performance of Linear, Transfer, and ReFine in which a model is pre-trained on miniIN or tieredIN and then fine-tuned on BSCD-FSL. In most cases, ReFine outperforms Linear and Transfer. This implies that random parameters are generally better than the source-fitted parameters, especially of the topmost layers, for fine-tuning initialization. Meanwhile, in the ISIC and ChestX data, we observed that it might be advantageous to transfer the source information to the target without re-randomization when the source data becomes larger.

4.3 Ablation Study on Where to Re-randomize

We demonstrate that re-randomizing the extractor at the topmost layers is essential. We only consider Transfer as a baseline for fair comparison because ReFine fine-tunes the entire network. Figure 3 shows the test accuracy according to the re-randomized stage(s).
We propose ReFine (re-randomization before fine-tuning), a simple yet effective method for CD-FSL, that involves resetting the parameters fitted to the source domain in order to maximize the efficacy of few-shot adaptation to the labeled target dataset. We demonstrate that our method outperforms conventional baselines under the CD-FSL setup. Furthermore, we investigate where and how to re-randomize the pre-trained models. We believe that our research will inspire CD-FSL researchers with the concept of re-initialization in the context of few-shot learning.

**Table 2: 5-way k-shot test accuracy over 600 tasks on [miniIN] \(\rightarrow\) [BSCD-FSL] according to the parts of re-randomization in the last stage. The topmost layers are boldfaced.**

| Path | Layer | Re-randomization layer |
|------|-------|------------------------|
| Original | Conv1 | ✓ ✓ ✓ ✓ ✓ |
|    | BN1   | ✓ ✓ ✓ ✓ ✓ |
|    | Conv2 | ✓ ✓ ✓ ✓ ✓ |
|    | BN2   | ✓ ✓ ✓ ✓ ✓ |
| Shortcut | ShortcutConv | ✓ ✓ ✓ ✓ ✓ |
|      | ShortcutBN | ✓ ✓ ✓ ✓ ✓ |

1-shot

| Dataset | 1-shot | 5-shot |
|---------|--------|--------|
| CropDisease | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| EuroSAT | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ISIC | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ChestX | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |

**Table 3: 5-way k-shot test accuracy over 600 tasks on [tieredIN] \(\rightarrow\) [BSCD-FSL] according to the parts of re-randomization in the last stage. The topmost layers are boldfaced.**

| Path | Layer | Re-randomization layer |
|------|-------|------------------------|
| Block1.Conv1 | ✓ ✓ ✓ ✓ ✓ |
| Block1.BN1 | ✓ ✓ ✓ ✓ ✓ |
| Block1.Conv2 | ✓ ✓ ✓ ✓ ✓ |
| Block1.BN2 | ✓ ✓ ✓ ✓ ✓ |
| Block1.Conv3 | ✓ ✓ ✓ ✓ ✓ |
| Block1.BN3 | ✓ ✓ ✓ ✓ ✓ |
| Block1.ShortCutConv | ✓ ✓ ✓ ✓ ✓ |
| Block2.Conv1 | ✓ ✓ ✓ ✓ ✓ |
| Block2.BN1 | ✓ ✓ ✓ ✓ ✓ |
| Block2.Conv2 | ✓ ✓ ✓ ✓ ✓ |
| Block2.BN2 | ✓ ✓ ✓ ✓ ✓ |
| Block2.Conv3 | ✓ ✓ ✓ ✓ ✓ |
| Block2.BN3 | ✓ ✓ ✓ ✓ ✓ |

1-shot

| Dataset | 1-shot | 5-shot |
|---------|--------|--------|
| CropDisease | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| EuroSAT | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ISIC | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ChestX | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |

5-shot

| Dataset | 1-shot | 5-shot |
|---------|--------|--------|
| CropDisease | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| EuroSAT | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ISIC | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |
| ChestX | ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ |

**Table 4: Analysis on the initializing distribution of ReFine. Sparse distribution initializes parameters with 20% sparsity. Lottery indicates re-initialization.**

| Shot | Distribution | CropDisease | EuroSAT | ISIC | ChestX |
|------|--------------|-------------|---------|------|--------|
| Uniform | 68.93 ± 0.84 | 64.14 ± 0.82 | 35.30 ± 0.59 | 22.48 ± 0.41 |
| Normal | 69.34 ± 0.86 | 60.85 ± 0.82 | 31.35 ± 0.58 | 22.38 ± 0.39 |
| 1 Orthogonal | 67.96 ± 0.84 | 59.71 ± 0.83 | 31.05 ± 0.59 | 22.50 ± 0.38 |
| Sparse | 69.07 ± 0.84 | 61.21 ± 0.82 | 31.10 ± 0.61 | 22.52 ± 0.39 |
| Lottery | 61.53 ± 0.92 | 61.30 ± 0.88 | 31.27 ± 0.57 | 21.87 ± 0.36 |

**4.4 Ablation Study on How to Re-randomize**

Table 4 shows that re-randomizing the parameters following uniform distribution is generally the best practice. Uniform and Normal indicate that the values are sampled from the uniform and normal distribution. Orthogonal indicates the weights are randomized as an orthogonal matrix, as described in Saxe et al. [30]. Sparse indicates the weights are randomized as a sparse matrix, where non-zero elements are sampled from the zero-mean normal distribution, as described in Martens et al. [20]. Lottery refers to re-initialization, i.e., resetting parameters to their initial state, prior to training. In the model pruning literature, the lottery ticket hypothesis [12] suggests that re-initialization can improve performance. However, we find that re-randomization is better suited in the case of CD-FSL. We believe that although re-initialization can be helpful in the original domain, this is not true under domain differences.

5 CONCLUSION

We propose ReFine (re-randomization before fine-tuning), a simple yet effective method for CD-FSL, that involves resetting the parameters fitted to the source domain in order to maximize the efficacy of few-shot adaptation to the labeled target dataset. We demonstrate that our method outperforms conventional baselines...
REFERENCES

[1] Ibrahim Abulaloumlgin, Hartmut Maen dell, and Daniel Keysers. 2021. The Impact of Reinitialization on Generalization in Convolutional Neural Networks. arXiv preprint arXiv:2109.00267 (2021).

[2] Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W Hoffman, David Pau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In Proceedings of the 30th International Conference on Neural Information Processing Systems. 3985–3996.

[3] Haoxing Chen, Huaxiong Li, Yaochu Li, and Chunlin Chen. 2021. Multi-level metric learning for few-shot image recognition. arXiv preprint arXiv:2103.11383 (2021).

[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In ICLR.

[5] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Durka, David Gutman, Brian Helba, Audi Kallio, Konstantinos Llopis, Michael Marchetti, et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019).

[6] Debasmit Das, Sungrack Yun, and Fatih Porikli. 2021. Cross-LE: A Framework for Single Source Cross-Domain Few-Shot Learning. In International Conference on Learning Representations.

[7] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2020. A Baseline for Few-Shot Image Classification. In ICLR.

[8] Rafael Rego Drumond, Lukas Brinkmeyer, Josif Grabocka, and Lars Schmidt-Thieme. 2020. HDRA: Head initialization across dynamic targets for robust architectures. In IJCAI. 397–405.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML. 1126–1135.

[10] Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. arXiv preprint arXiv:1806.02837 (2018).

[11] Sebastian Flennerhag, Andrei A. Rusa, Ravzan Pascaaru, Francesco Visin, Hujun Yin, and Raia Hadsell. 2020. Meta-Learning with Warped Gradient Descent. In International Conference on Learning Representations. https://openreview.net/forum?id=rKQBFPP.

[12] Jonathan Frankle and Michael Carbin. 2018. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In International Conference on Learning Representations.

[13] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. 2020. A broader study of cross-domain few-shot learning. In ECCV. 124–141.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In CVPR.

[15] Patrick Helber, Benjamin Bischoke, Andreas Dangel, and Damian Both. 2019. EuroSat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 7 (2019), 2217–2226.

[16] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. 2021. Dynamic Distillation Network for Cross-Domain Few-Shot Recognition with Unlabeled Data. arXiv preprint arXiv:2012.06707 (2021).

[17] Xingjian Li, Haoyi Xiong, Haozhe An, Cheng-Zhong Xu, and Dejing Dou. 2020. ReFine: Re-randomization before Fine-tuning for Cross-domain Few-shot Learning. In CIKM ’22, October 17–21, 2022, Atlanta, GA, USA.

[18] Hanwen Liang, Qiong Zhang, Peng Dai, and Juewu Lu. 2021. Boosting the Generalization Capability in Cross-Domain Few-Shot Learning via Noise-enhanced Supervised Autoencoder. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9424–9434.

[19] Hartmut Maennel, Ibrahim Abulaloumlgin, Ilya Tolstikhin, Robert JN Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. 2020. What Do Neural Networks Learn When Trained With Random Labels? arXiv preprint arXiv:2006.10455 (2020).

[20] James Martens et al. 2010. Deep learning via hessian-free optimization. In ICML. Vol. 27. 735–742.

[21] Sharan Arora, Yao Sun, Ilya Sutskever, and战胜. 2016. Using deep learning for image-based plant disease detection. Frontiers in plant science (2016), 1419.

[22] Jaehee Oh, Sungnyun Kim, Namgyu Ho, Jon-Hwa Kim, Hwanjun Song, and Se-Young Yun. 2022. Understanding Cross-Domain Few-Shot Learning: An Experimental Study. arXiv preprint arXiv:2202.01339 (2022).

[23] Jaehee Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. 2021. BOIL: Towards Representation Change for Few-Shot Learning. In International Conference on Learning Representations. https://openreview.net/forum?id=umJiUL5qMH.

[24] Cheng Peng, Phoo and Bharath Hariharan. 2021. Self-Tuning For Few-Shot Transfer Across Extreme Task Differences. In ICLR.

[25] Anuruddha Raghu, Mathira Raghu, Samy Bengio, and Oriol Vinyals. 2019. Rapid learning or feature reuse? towards understanding the effectiveness of maml. arXiv preprint arXiv:1909.09157 (2019).

[26] Mathira Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208 (2019).

[27] Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W Hoffman, David Pau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In Proceedings of the 30th International Conference on Neural Information Processing Systems. 3985–3996.

[28] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, and Richard S Zemel. 2018. Meta-learning. arXiv preprint arXiv:1803.06706 (2018).

[29] Youngmin Ro, Jongwon Choi, Dae Ung Jo, Byeongho Heo, Jongin Lim, and Jin Young Choi. 2019. Backbone cannot be trained at once: Rolling back to pre-trained network for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 8859–8867.

[30] Andrew M. Saxe, James McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 (2013).

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

[32] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In CVPR. 1199–1208.

[33] Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. arXiv preprint arXiv:2004.14975 (2020).

[34] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. 2020. Rethinking few-shot image classification: a good embedding is all you need? In ECCV. 266–282.

[35] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are neural network models? In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 8859–8867.

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one-shot learning. NeurIPS 29 (2016), 3630–3638.

[37] Xiaosong Yang, Wifeng Pang, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In CVPR. 2097–2106.

[38] Yaqing Wang, Qianbing Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR) 53, 3 (2020), 1–34.

[39] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? arXiv preprint arXiv:1412.1794 (2014).

[40] Chiyuan Zhang, Samy Bengio, and Yoram Singer. 2019. Are all layers created equal? arXiv preprint arXiv:1902.01996 (2019).

[41] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2022. Experimental Study. arXiv preprint arXiv:2004.14975 (2020).

[42] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon White. 2019. Fast context adaptation via meta-learning. In ICML. 7693–7702.