Uniform Priors for Data-Efficient Learning

Samarth Sinha[1,*], Karsten Roth[2,*], Anirudh Goyal[3], Marzyeh Ghassemi[4],
Zeynep Akata[2,5], Hugo Larochelle[3,6], Animesh Garg[1,7,8]
[1]University of Toronto, [2]University of Tübingen, [3] MILA, [4]MIT, [5]MPI-IS,
[6]Google, [7]Vector Institute, [8]Nvidia

Abstract

Few or zero-shot adaptation to novel tasks is important for the scalability and deployment of machine learning models. It is therefore crucial to find properties that encourage more transferable features in deep networks for generalization. In this paper, we show that models that learn uniformly distributed features from the training data, are able to perform better transfer learning at test-time. Motivated by this, we evaluate our method: uniformity regularization ($\mathcal{UR}$) on its ability to facilitate adaptation to unseen tasks and data on six distinct domains: Few-Learning with Images, Few-shot Learning with Language, Deep Metric Learning, 0-Shot Domain Adaptation, Out-of-Distribution classification, and Neural Radiance Fields. Across all experiments, we show that using $\mathcal{UR}$, we are able to learn robust vision systems which consistently offer benefits over baselines trained without uniformity regularization and are able to achieve state-of-the-art performance in Deep Metric Learning, Few-shot learning with images and language.

1. Introduction

Deep Neural Networks have enabled great success in various machine learning domains such as computer vision [14, 19, 38] and , natural language processing [3, 9, 68]. Despite their success, they still suffer from not adapting well to a novel data distribution that they have not previously encountered during training time due to a *distribution shift*. This motivates the problem of learning more transferable features at training time, that can then adapt to a novel data distribution that the model encounters at test-time.

Understanding how to achieve generalization under such distributions shifts is an active area of research. In the few-shot Meta-Learning setting [6, 11, 57], a meta-learner is tasked to quickly adapt to novel test data given its training experience and a limited labeled data budget. Similarly, in Deep Metric Learning (DML) [16, 51] and Zero-Shot Domain Adaptation (ZSDA), [30, 67] study generalization at the limit of such adaptation, where predictions on novel test data are made without any test-time finetuning. Yet, despite the motivational differences, each of these fields require representations to be learned from the training data that allow for better generalization to novel tasks and data. Although there exists a large corpus of domain-specific training methods, in this paper we seek to determine fundamental properties that learned features and feature spaces should have to facilitate such generalization.

Fortunately, recent literature provides pointers towards one such property: the notion of "feature uniformity" for improved generalization. Feature uniformity suggests that if we are able to learn representations that are more uniformly distributed at training time, then the model is able to generalize better to a novel data distribution. For unsupervised representation learning, [70] highlight a link between the uniform distribution of hyperspherical feature representations and the transfer performance in downstream tasks, which has been implicitly adapted in the design of modern contrastive learning methods [2, 62, 63].

Similarly, [51] show that for Deep Metric Learning, uniformity in coverage and uniform singular value distribution of the learned embedding spaces are strongly connected to zero-shot generalization performance. Both [70] and [51] link uniformity in the feature space to preservation of maximal "reusable" information for zero-shot generalization. This suggests that actively imposing a uniformity prior on learned feature representations should encourage better transfer properties by retaining more information and reducing bias towards training tasks, which in turn facilitate better adaptation to novel tasks at test time.

However, while both [70] and [51] propose methods to incorporate this notion of uniformity, they are defined only for hyperspherical embedding spaces or contrastive learning approaches[1], thus severely limiting the applicability to other domains such as supervised learning and meta-learning.

To address these limitations and leverage the benefits of uniformity for all novel task and data adaptation for deep neural networks, we propose *uniformity regulariza-*

[1]By imposing a Gaussian potential over hyperspherical embedding distances or pairwise sample relations.

tion, which places a uniform hypercube prior on the learned feature space during training, without being limited to the contrastive training approaches or a hyperspherical representation space. Unlike e.g. a multivariate Gaussian, the *uniform* prior puts equal likelihood over the feature space, which then enables the network to make fewer assumptions about the data, limiting model overfitting to the training task. Our *uniformity regularization* follows adversarial learning frameworks that allow us to apply our proposed uniformity prior over features without the need to assume an explicit parameterization for the network feature distribution to define a closed-form KL-Divergence objective.

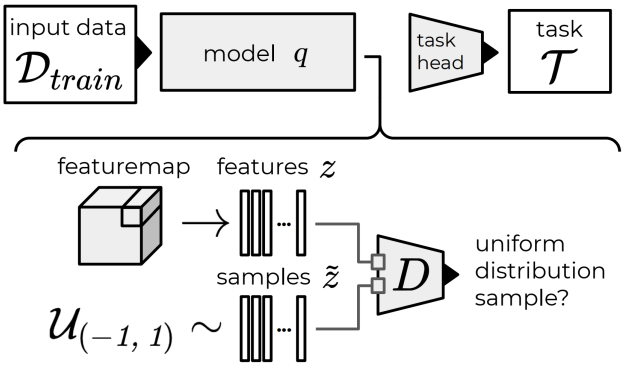
Using this setup, we experimentally demonstrate that *uniformity regularization* aids generalization in zero-shot setups such as Deep Metric Learning, Domain Adaptation, Out-of-Distribution Detection, few-shot learning for both vision and language, and neural radiance fields. Furthermore, for Deep Metric learning and few-shot learning, we are able to set a new state-of-the-art over large-scale datasets [31, 66, 73].

2. Related Work

Adversarial Representation Learning. Latent variable models (e.g. [39, 65]) have used GAN-style training in the latent space to learn a rich posterior. Recent efforts have made such training effective in different contexts like active learning [26, 56] or domain adaptation [22, 67]. Our work more closely follows [21], who applied a similar adversarial objective to impose certain properties on the features learned through self-supervised representation learning. While [21] also matched their representations to a uniform distribution, no detailed reasons for this specific choice were given. Instead, our work shows that it is exactly the uniformity of feature distributions introduced by our *uniformity regularization* that facilitate fast adaptation and transfer to novel data and tasks in neural networks, regardless of the specific application domain.

Deep Metric Learning and Generalization. The goal of a Deep Metric Learning (DML) algorithm is to learn a metric space that encodes semantic relations as distances and which generalizes sufficiently that at test-time zero-shot retrieval on novel classes and samples can be performed. Representative methods in DML commonly differ in their proposed objectives [5, 16, 71, 74, 79], which are commonly accompanied with tuple sample methods [18, 50, 72, 74]. Extension to the basic training paradigm, such as with self-supervision [4, 41] have also shown great promise. Recently, [51] performed an extensive survey on various DML objectives to study driving factors for generalization among these methods. In that regard, recent work by [70] has offered theoretical insights into the benefits of learning on a Uniform hypersphere for zero-shot generalization.
Achieving Out-of-Distribution generalization from different point of views has also been of great interest, ranging from work on zero-shot domain adaptation [22, 30, 67] to the study of invariant correlations [1].

Figure 1. **Incorporating Uniform Priors.** The network is tasked to adversarially learn features that fool a discriminator trained to classify samples drawn a uniform distribution, allowing us to impose a uniform distribution prior over the feature distribution without requiring a closed-form KL-Divergence.



Meta-Learning Many types of meta-learning algorithms for few-shot learning (but also for zero-shot learning such as [44]) have recently been proposed, building on memory-augmented methods [43, 47, 54], metric-based approaches [57, 59, 69] or optimization-based techniques [11, 36, 46, 77]. Finetuning using ImageNet pretraining [6, 13] has also been proposed as alternative approaches. Meta-learning has also been explored for fast adaptation of novel tasks in reinforcement learning [24, 29, 81]. More closely related to our approach is [25], proposing inequality measures between different tasks for less task-dependent representations. However, this is still limited to episodic learning akin to most few-shot learning approaches. *Uniformity regularization* is much more generic, being applicable to domains outside of Meta-Learning, and does not dependent on the choice of inequality measure. Fast adaptation has recently been popularized by different meta-learning strategies [11, 57]. These methods assume distinct meta-training and meta-testing task distributions, where the goal of a meta-learner is to adapt fast to a novel task given limited samples for learning it.

3. Background

Generative Adversarial Networks (GANs) Generative Adversarial Networks (GANs) were proposed to optimize a min-max game between a generator G and a discriminator D. The generator $G(z)$ is trained to map samples from a prior $z \sim p(z)$ to the target space, while the discriminator is trained to be an arbiter between the target data distribution $p(x)$ and the generator distribution. The training objective

can be written as:

$$\mathcal{L}_D = \max_D \mathbb{E}_{z\sim p(z)}[1 - \log D(G(z))] + \mathbb{E}_{x\sim p(x)}[\log D(x)] \quad (1)$$

$$\mathcal{L}_G = \min_G \mathbb{E}_{z\sim p(z)}[1 - \log D(G(z))] \quad (2)$$

with $p(z)$ the generator prior and $p(x)$ a defined target distribution (e.g. natural images).

Generalization to novel data While meta-learning approaches assumes the availability of a finetuning budget for adaptation at test time, zero-shot approaches introduce the limit scenario of fast adaptation, in which generalization has to be achieved without access to any examples. Such a setting can be found in metric learning [60, 76], where a model is evaluated on the ability to perform zero-shot retrieval on novel data. Most commonly, metric models are trained on a training data distribution $\mathcal{D}_{\text{train}}$ and evaluated on a testing distribution $\mathcal{D}_{\text{test}}$ which share no classes. However, the data generating function is assumed to be similar between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, such as natural images of birds [73]. Similar to DML, Zero-Shot Domain Adaptation (ZSDA) introduces a learner that is also trained and evaluated on two distinct $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ in a zero-shot setting. However, unlike DML, in ZSDA, the labels between the data distributions are shared. Instead, training and test distribution come from distinct data generative functions, such as natural images of digits [15] and handwritten images of digits [34].

4. Training with Uniformity Regularization

In this section, we introduce the proposed *uniformity regularization* and detail the employed alternating GAN-like optimization scheme to perform it in a computationally tractable manner.

Prior Matching. Given a neural network $q(y|x)$ that is parameterized by θ we formally define the training objective as $\mathcal{L}_T(q(y|x), y)$ where $\mathcal{L}_T$ is any task-specific loss such as a cross-entropy loss, (x, y) are samples from the training distribution $\mathcal{D}_{\text{train}}$ and $q(y|x)$ is the probability of predicting label y under q. This is a simplified formulation; in practice, there are many different ways to train a neural network, such as ranking-based training with tuples [7]. We define the embedding space z as the output of the final convolutional layer of a deep network. Accordingly, denote the conditional distribution for that embedding space by $q(z|x)$ which, due to the neural network being a deterministic mapping, is a Dirac delta distribution at the value of the final convolutional layer. Section 5.1 further details how to apply *uniformity regularization* in practice.

As we ultimately seek to impose a uniformity prior over the learned aggregate feature/embedding "posterior" $q(z) = \int_x q(z|x)p(x)dx$, we begin by augmenting the generic task-objective to allow for the placement of a prior $r(z)$. For priors $r(z)$ with closed-form KL-divergences $\mathbf{D}^{\text{KL}}$ and a parametrization for the feature distribution $q(z)$ (e.g. by assuming $z \sim \mathcal{N}(\mu, \sigma)$), one can define a prior-regularized task objective as

$$\mathcal{L} = \min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_T(q(y|x), y)\right] + \mathbf{D}^{\text{KL}}_{x\sim\mathcal{D}_{\text{train}}}\left(q(z|x)\|r(z)\right), \quad (3)$$

similar to the Variational Autoencoder formulation in [28]. However, this requires us to have an explicit formulation for the distribution of embeddings, which severely impacts the general applicability while requiring prior knowledge about the expected feature distribution. In addition to that, defining a KL-Divergence of the form $\mathbf{D}^{\text{KL}}_{x\sim\mathcal{D}_{\text{train}}}(q(z|x)\|r(z))$ where $r(z)$ is our uniform distribution $\mathcal{U}(-\alpha, \beta)$ with lower and upper bounds $-\alpha$ and β, is not well-defined outside of $[-\alpha, \beta]$ (where $\mathcal{U}(-\alpha, \beta)$ is zero), while $q(z|x)$ can have non-zero probability density in that region.

Uniformity Regularization. To address the practical limitation of solving Eqn. 3 with a uniform prior and without constraining potential feature distributions $q(z|x)$, we draw upon the GAN literature (as briefly introduced in section 3, in which alternate adversarial optimization has been successfully used to match a generated to a predefined target distribution using implicit divergence minimization. Latent variable models such as the Adversarial Autoencoder [39] have successfully used such a GAN-style adversarial loss instead of a KL divergence in the latent space of the autoencoder to learn a rich posterior. Such implicit divergence minimization allows us to match any well-defined distribution as a prior, but more specifically, ensures that we can successfully match learned embedding spaces to $\mathcal{U}(-\alpha, \beta)$, which we set to the unit hypercube $\mathcal{U}(-1, 1)$ by default.

To this end, we adapt the GAN objective in Eqns. 1 and 2 for uniformity regularization optimization and train a discriminator, D, to be an arbiter between which samples are from the learned distribution $q(z|x)$ and from the uniform prior $r(z)$. As such, the task model q (parameterized by θ) aims to *fool* the discriminator D into thinking that learned features, $q(z|x)$, come from the chosen uniform target distribution, $r(z)$, while the discriminator D learns to distinguish between learned features and samples taken from the prior, $\tilde{z} \sim r(z)$. Note that while the task-model defines a deterministic mapping for $q(z|x)$ instead of a stochastic one, the aggregate feature "posterior" $\int_x q(z|x)p(x)dx$, on which we apply our uniformity prior, is indeed a stochastic distribution [39].

Concretely for our *uniformity regularization*, we rewrite the discriminator objective from Eqn. 1 to account for the

Table 1. **Influence of Feature Space Uniformity on Generalization.** We study the influence of feature space uniformity on generalization in ZSDA by matching the feature space to prior distributions $r(z)$ of increasing uniformity (left to right). We report mean accuracy and standard deviation over 5 runs on the task of MNIST → USPS and USPS → MNIST zero-shot domain adaptation using ResNet-18.

Task	Baseline	$\mathcal{N}(0, 0.1 \times \mathcal{I})$	$\mathcal{N}(0, \mathcal{I})$	$\mathcal{N}(0, 5 \times \mathcal{I})$	$\mathcal{N}(0, 10 \times \mathcal{I})$	$\mathcal{N}(0, 25 \times \mathcal{I})$	$\mathcal{U}(-1, 1)$
MNIST → USPS	49.0 ± 0.20	43.98 ± 0.23	43.45 ± 0.16	56.45 ± 0.36	59.80 ± 0.12	50.11 ± 0.23	**67.2** ± 0.11
USPS → MNIST	42.8 ± 0.07	27.23 ± 0.28	26.02 ± 0.87	37.96 ± 0.32	43.76 ± 0.48	32.90 ± 0.45	**56.2** ± 0.10

uniform prior matching, giving

$$\mathcal{L}_D = \max_D \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}}[\log(1 - D(q(z|x)))] + \mathbb{E}_{\tilde{z} \sim \mathcal{U}(-1,1)}[\log D(\tilde{z})]. \tag{4}$$

Consequently, we reformulate the generator objective from Eqn. 2 to reflect the task-model q,

$$\mathcal{L}_{\text{max}} = \min_\theta \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}}[\log(1 - D(q(z|x)))], \tag{5}$$

where we used the notation $\mathcal{L}_{\text{max}}$ to reflect that the optimization maximizes the feature uniformity by learning to fool D. Our final min-max *uniformity regularized* objective for θ and the Discriminator is then given as

$$\mathcal{L} = \min_\theta \max_D \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{train}}}[\mathcal{L}_T(q_\theta(y|x), y)] + \gamma \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}}[\log(1 - D(q_\theta(z|x)))] + \mathbb{E}_{\tilde{z} \sim \mathcal{U}(-1,1)}[\log D(\tilde{z})] \tag{6}$$

with task-objective $\mathcal{L}_T$ and training data distribution $\mathcal{D}_{\text{train}}$. Using this objective, the learned feature space is implicitly encouraged to become more uniformly distributed. The amount of regularization is controlled by the hyperparameter γ, balancing generalization of the model to new tasks and performance on the training task at hand. Large γ values hinder effective feature learning from training data, while values of γ that are too small result in weak regularization, leading to a non-uniform learned feature distribution with reduced generalization capabilities.

5. Experiments

We begin by highlighting the link between feature space uniformity and generalization performance (§5.2). In a large-scale experimental study covering settings in which samples are available for adaptation (Meta-Learning, §5.3), or not (Deep Metric Learning & Zero-Shot Domain Adaptation, §5.4) and Out-of-Distribution Detection (§5.5), we then experimentally showcase how *uniformity regularization* can facilitate generalizability of learned features and the ability of a model to perform fast adaptation to novel tasks and data.

For all experiments, *no hyperparameter tuning on base algorithms is done*, and the same hyperparameters that the respective original papers proposed are used; we simply add the *uniformity regularization*, along with the task loss as in Eqn. 6.

5.1. Experimental Details

Uniformity regularization was added to the output of the CNNs for all networks. For ResNet-variants [19, 75, 78], it was applied to the output of the CNNs, just before the single fully-connected layer. For meta-learning, the regularization is applied directly on the learned metric space for the metric-space based meta-learners [37, 57, 69], and applied to the output of the penultimate layer for MAML [11]. The discriminator is parameterized using a three-layer MLP with 100 hidden units in each layer and trained using the Adam optimizer [27] with a learning rate of 10^{-5}. The value of γ is chosen to be 0.1 for all experiments, except for Deep Metric Learning. For Deep Metric Learning, a value of $\gamma = 0.4$ is chosen, since the effect of regularization needs to be stronger, as Deep Metric Learners (commonly a ResNet-50 [19] or Inception-V1 [61] with BatchNorm [23]) start off with networks that are already pretrained on ImageNet [52].

5.2. Feature Space Uniformity is linked to Generalization Performance

We first investigate the connection between feature space uniformity and generalization, measured by generalization performance in Zero-Shot Domain Adaptation (more experimental details in §5.4). Unfortunately, the uniform hypercube prior in our *uniformity regularizer* does not provide a way for intuitive and explicit uniformity scaling - one can not make the uniform prior "more or less uniform".

As such, we make use of a Gaussian prior $\mathcal{N}(\mu, \sigma^2)$. Under the fair assumption that the learned embedding space of deep neural networks does not have infinite support in practice (especially given regularization methods such as L2 regularization), the variance σ^2 provides a uniformity scaling factor - with increased variance, the Gaussian prior reduces mass placed around embeddings near μ, effectively encourageing the network to learn a more uniform embedding space. We can therefore directly evaluate the importance of feature space uniformity by using our GAN-based regularization scheme to match feature space distribution to Gaussian priors with different σ^2 scales.

Table 1 compares feature space uniformity against the model's ability to perform ZSDA from MNIST to USPS (and respective backward direction) using a ResNet-18 [19] (with "*Baseline*" the unregularized model). As can be seen,

Table 2. **Meta-Learning**. **1)** Comparing with the state of the art wrt different regularization techniques, e.g. Dropout, L2, vs our uniformity alignment on Omniglot, Double MNIST, CIFAR-FS and miniIMageNet with the same setting as [70].We report the mean **error rate** for Omniglot & Double MNIST and mean **accuracies** for CIFAR-FS and miniImageNet over 5 seeds. We use the exact hyperparameters as proposed in the original paper of each meta-learner. **2)** Applying our UR with Universal Representation Transformer Layers [37] on the Meta-Dataset to establish the new state-of-the-art (**blue**).

1) Baseline Study	Omniglot		Double MNIST		CIFAR-FS		miniImageNet	
Methods ↓	(5, 1)	(5,5)	(5, 1)	(5,5)	(5, 1)	(5,5)	(5, 1)	(5,5)
MAML	**4.8**± 0.4	**1.5** ± 0.4	**7.9**± 0.7	**1.9**± 0.3	**52.1** ± 0.8	**67.1** ± 0.9	47.2 ± 0.7	**62.1** ± 1.0
MAML + $\mathcal{UR}$	**4.1**± 0.5	**1.3**± 0.2	**7.3** ± 0.2	**1.5** ± 0.5	**52.9** ± 0.4	**67.1** ± 0.9	**48.9** ± 0.8	**64.1** ± 1.0
Matching Networks	2.1 ± 0.2	**1.0** ± 0.2	4.2 ± 0.2	**2.7** ± 0.2	46.7 ± 1.1	**62.9** ± 1.0	43.2 ± 0.3	50.3 ± 0.9
Matching Networks + Dropout	2.4 ± 0.2	1.3 ± 0.2	4.4 ± 0.2	2.9 ± 0.4	45.3 ± 1.1	**63.0** ± 0.7	42.9 ± 0.9	50.0 ± 1.0
Matching Networks + L2 reg.	2.1 ± 0.2	**1.0** ± 0.1	4.1 ± 0.2	**2.6** ± 0.2	46.9 ± 1.1	**63.0** ± 0.9	43.3 ± 0.8	50.1 ± 1.0
Matching Networks + U-A	2.0 ± 0.1	**0.9** ± 0.1	3.9 ± 0.3	**2.7** ± 0.1	47.3 ± 1.0	**63.1** ± 0.8	43.5 ± 0.7	50.3 ± 1.0
Matching Networks + $\mathcal{UR}$	**1.7**± 0.1	**0.9**± 0.1	**3.2**± 0.1	**2.3**± 0.3	**49.3** ± 0.4	**63.1** ± 0.7	**47.1** ± 0.8	**53.1** ± 0.7
Prototypical Network	**1.6** ± 0.2	**0.4** ± 0.1	**1.3** ± 0.2	**0.2** ± 0.2	**52.4** ± 0.7	**67.1** ± 0.5	45.4 ± 0.6	61.3 ± 0.7
Prototypical Network + Dropout	1.9 ± 0.2	**0.5** ± 0.2	**1.4** ± 0.2	0.5 ± 0.1	51.9 ± 0.8	**66.0** ± 0.4	44.8 ± 0.7	61.2 ± 0.9
Prototypical Network + L2 reg.	**1.6** ± 0.2	**0.4** ± 0.1	**1.3** ± 0.1	**0.3** ± 0.2	**52.5** ± 0.8	**66.3** ± 0.4	45.0 ± 0.7	61.4 ± 0.7
Prototypical Network + U-A	**1.5** ± 0.3	**0.4** ± 0.1	**1.2** ± 0.1	**0.2** ± 0.2	**52.6** ± 0.7	**66.3** ± 0.5	45.4 ± 0.5	61.8 ± 0.8
Prototypical Network + $\mathcal{UR}$	**1.2** ± 0.3	**0.4** ± 0.1	**1.0** ± 0.2	**0.2** ± 0.2	**52.6** ± 0.8	**66.8** ± 0.5	**46.8** ± 0.5	**64.4** ± 0.9

2) Meta-Dataset	ILSVRC	Omniglot	Aircrafts	Birds	Textures	QuickDraw	Fungi	VGGFlower	TrafficSigns	MSCOCO	Avg. Rank
TaskNorm	50.6 ± 1.1	90.7 ± 0.6	83.8 ± 0.6	74.6 ± 0.8	62.1 ± 0.7	74.8 ± 0.7	48.7 ± 1.0	89.6 ± 0.6	67.0 ± 0.7	43.4 ± 1.0	4.5
SUR	56.3 ± 1.1	93.1 ± 0.5	85.4 ± 0.7	71.4 ± 1.0	71.5 ± 0.8	81.3 ± 0.8	**63.1** ± 1.0	82.8 ± 0.7	70.4 ± 0.8	52.4 ± 1.1	3.2
SimpleCNAPS	**58.6** ± 1.1	91.7 ± 0.6	82.4 ± 0.7	74.9 ± 0.8	67.8 ± 0.8	77.7 ± 0.7	46.9 ± 1.0	**90.7** ± 0.5	**73.5** ± 0.7	46.2 ± 1.1	3.2
URT	55.7 ± 1.0	94.4 ± 0.4	85.8 ± 0.6	**76.3** ± 0.8	71.8 ± 0.7	82.5 ± 0.6	**63.5** ± 1.0	88.2 ± 0.6	69.4 ± 0.8	52.2 ± 1.1	2.6
URT + $\mathcal{UR}$	**58.3** ± 0.9	**95.2** ± 0.2	**88.0** ± 0.9	**76.7** ± 0.8	**74.9** ± 0.9	**84.0** ± 0.3	**62.8** ± 1.1	**90.3** ± 0.4	**72.9** ± 0.8	**54.6** ± 1.1	**1.5**

when the uniformity of the (Gaussian) prior $r(z)$ is increased up to a certain breaking point, the ability to perform domain adaptation also improves. When σ^2 is small, the model is unable to effectively adapt to the novel data, and as the uniformity of $r(z)$ is increased, the network significantly improves its ability to perform the adaptation task. However, for very large sigma, value scales become an issue, and training becomes less stable, with performance dropping. This motivates the use of our maximally uniform hyper-cube prior $\mathcal{U}(-1, 1)$, which significantly improves the performance, coinciding with insights made in [70] and [51].

The impact of our *uniformity regularization* is even more evident on the backward task of USPS → MNIST, since there are less labels present in the USPS dataset, thereby making overfitting a greater issue when trained on USPS.

5.3. Uniform Priors benefit Meta-Learning

We now study the influence of *uniformity regularization* on meta-training for few-shot learning tasks, which we divide into two experiments.

First, we evaluate how *uniformity regularization* impacts the performance of three distinct meta-learning baselines: Matching Networks [69], Prototypical Networks [57] and MAML [11]. Performance is evaluated on four few-shot learning benchmarks: Double MNIST [34], Omniglot [33], CIFAR-FS [32] and miniImagenet [69]. For our implementation, we utilize TorchMeta [8]. Results for each meta-learning method with and without regularization are summarized in Table 2a)[2]. We observe that, adding *uniformity regularization* benefits generalization across method and benchmark, in some cases notably. This holds regardless of the number of shots used at meta-test-time, though we find the largest performance gains in the 1-shot scenario.

In addition, when compared to other regularization methods such as Dropout [58], L2-regularization [64] and hyperspherical uniformity regularization [70], it compares favorably, especially on more complex datasets such as miniImageNet. Compared to [70], this is especially impressive given the much wider application range. Overall, the results highlight the benefit of reduced training-task bias introduced by *uniformity regularization* for fast adaptation to novel test tasks.

Second, we examine *uniformity regularization* on the Meta-Dataset [66], which contains data from diverse do-

[2]For Double MNIST and Omniglot, we list error rates, not accuracies.

Table 3. **Few-Shot Language Relation Classification** using transformers (BERT pretrained) on non-image data.

Method ↓	5-way 1-shot	Method ↓	5-way 1-shot
Bert-PAIR	85.4 ± 0.3	**Bert-Proto**	78.7 ± 0.8
$+\mathcal{UR}$	86.4 ± 0.2	$+\mathcal{UR}$	80.6 ± 0.6

mains such as natural images, objects and drawn characters. We follow the setup suggested by [66], used in [37], in which eight out of the ten available datasets are used for training, while evaluation is done over all. Results are averaged across varying numbers of ways and shots. We apply *uniformity regularization* on the state-of-the-art Universal Representation Transformer (URT) [37], following their implementation and setup without hyperparameter tuning. Table 2b), *uniformity regularization* shows consistent improvements upon URT, matching or even outperforming the state of the art on all sub-datasets.

To highlight that a uniformity prior does not only benefit specific convolutional architectures in the vision domain, we also evaluate few-shot generalization capacities in the language domain. Specifically, we select the complex task of fewshot relation classification as proposed in [17]. Here, given some support object relations (taken from [17]) such as "*London is the the capital of the U.K.*" or "*Newton served as the president of the Royal Society*" as well as their respective relation class (in this case "*capital_of*" and "*member_of*", respectively), the goal is then to classify unseen query relations (s.a. "*Euler was elected a foreign member of the Royal Swedish Academy of Sciences*" being a "*member_of*"-relation). As reference methods, we select BERT-Proto and BERT-PAIR introduced in [12], with BERT-PAIR providing competitive, near state-of-the-art performance. Both of these methods leverage a BERT-transformer-base [9, 68], on top of which the respective few-shot models *Prototypical Networks* and *PAIR* are applied.

Results are reported in Tab. 3 following the official dataset and code provided in [17], with scores representing performance on the official validation set, as the official test set is submission-locked. For training, we thus used a $85\% - 15\%$ training-validation split of the original training dataset. As can be seen, a consistent improvement in the most challenging 5-way 1-shot setting can seen, showcasing a general applicability of uniform priors beyond just image-data and networks.

5.4. Uniform Priors for Zero-Shot Generalization

Going further, we study limit cases of fast adaption and look at how *uniformity regularization* affects zero-shot retrieval in Deep Metric Learning and zero-shot classification for domain adaptation. Here, the model is evaluated on a different distribution than the training distribution without finetuning, highlighting the benefits of *uniformity regularization* for learning task-agnostic & reusable features.

Table 4. **Deep Metric Learning (Zero-Shot Generalization).** **1)** Evaluating our $\mathcal{UR}$ wrt state-of-the-art DML methods as in [51] with a ResNet-50 backbone [19] on CUB200-2011 [73] & CARS196 [31] datasets. We report Recall@1 and Normalized Mutual Information (NMI). **2)** With standard learning rate scheduling, our $\mathcal{UR}$ boosts performance of baseline objectives in the DML literature corpus.For table **2)**, numbers in **bold** represent the best performance for a given benchmark setting and metric.

1) Ablations	CUB200-2011		CARS196	
Methods ↓	R@1	NMI	R@1	NMI
ResNet50, Embedding Dim.: 128				
Softmax [79]	61.7 ± 0.3	66.8 ± 0.4	78.9 ± 0.3	66.4 ± 0.3
Softmax + $\mathcal{UR}$	**65.0** ± 0.1	**68.8** ± 0.2	**80.6** ± 0.2	**68.3** ± 0.2
Margin [74]	63.1 ± 0.5	68.2 ± 0.3	79.9 ± 0.3	67.4 ± 0.3
Margin + $\mathcal{UR}$	**65.0** ± 0.3	**69.5** ± 0.2	**82.5** ± 0.1	**68.9** ± 0.2
Msim [71]	62.8 ± 0.7	68.6 ± 0.4	81.7 ± 0.2	69.4 ± 0.4
Msim + $\mathcal{UR}$	**65.4** ± 0.4	**70.3** ± 0.3	**82.2** ± 0.2	**70.5** ± 0.3

2) Literature	CUB200-2011		CARS196	
Methods ↓	R@1	NMI	R@1	NMI
ResNet50, Embedding Dim.: 128				
Div&Conq [53]	65.9	69.6	**84.6**	70.3
MIC [49]	66.1	69.7	82.6	68.4
PADS [50]	**67.3**	69.9	83.5	68.8
Msim+$\mathcal{UR}$	66.3 ± 0.4	**70.5** ± 0.3	84.0 ± 0.2	**71.3** ± 0.5
Inception-V1 + BatchNorm, Embedding Dim.: 512				
Msim [71]	65.7	-	84.1	-
Softtriple [45]	65.4	69.3	84.5	70.1
Group [10]	65.5	69.0	**85.6**	**72.7**
Msim+$\mathcal{UR}$	**68.5** ± 0.3	**71.7** ± 0.5	**85.8** ± 0.3	**72.2** ± 0.5

Deep Metric Learning. We apply *uniformity regularization* on four benchmark DML objectives (Contrastive Loss [16], Margin Loss [74], Softmax Loss [79] and Multi-Similarity Loss [71]) studied in [51], and evaluate them over two standard datasets: CUB-200 [73], and Cars-196 [31]. The results summarized in Table 4a) reveal substantial gains over all evaluation metrics and benchmarks and a diverse set of baseline, for example for more than 3% when applied to Softmax Loss [79] on CUB200-2011 [73]. But even stronger baselines s.a. MultiSimilarity [71] see substantial gains on both datasets across all evaluation metrics, This showcases the effectiveness in fighting against overfitting for generalization, even without finetuning at test-time. Finally, when evaluated in two different common literature settings and compared against recent methods, we find that simple *uniformity regularized* objectives can match or even outperform these, in some cases significantly, as seen e.g. when applied jointly with the MultiSimilarity Loss

Table 5. **Zero-Shot Domain Adaptation**. Comparing several zero-shot domain adaptation strategies on the digit recognition task ("*From*/*To*") w and w/o $\mathcal{UR}$ reporting mean accuracy and std over 5 random seeds. The results for ADDA and the "Source Only" + LeNet (LN) backbone are taken directly from [67]. "Target Only" refers to direct training and evaluation on the target distribution. We perform no hyperparameter tuning, and the exact hyperparameters are used as in [67].

Method↓	MNIST/USPS	USPS/MNIST	SVHN/MNIST
Source Only (R18)	49.0 ± 0.20	42.8 ± 0.07	69.7 ± 0.06
+ $\mathcal{UR}$ (R18)	**67.2** ± 0.11	**56.2** ± 0.10	**71.3** ± 0.13
Source Only (LN)	75.2 ± 0.02	57.1 ± 0.02	60.1 ± 0.01
+ $\mathcal{UR}$ (LN)	**79.6** ± 0.04	**62.6** ± 0.01	**65.8** ± 0.03
ADDA [67] (LN)	89.4 ± 0.01	90.1±0.01	76.0 ± 0.02
+ $\mathcal{UR}$ (LN)	**93.5** ± 0.09	**94.8** ± 0.03	**81.6** ± 0.03
Target Only (R18)	98.1 ± 0.2	99.8 ± 0.1	99.8 ± 0.1

on CUB200-2011, where we e.g. beat multi-proxy SoftTriple [45] by more than 3%.

Zero-Shot Domain Adaptation. For Zero-Shot Domain Adaptation, we conduct digit recognition experiments, transferring models between MNIST [34], SVHN [15] and USPS [55]. In this setting, we train the model on a source dataset, and test it directly on the test dataset. Since each of the datasets contain digits, the networks are assessed on their ability to classify digits on the target dataset, without any training. We evaluate different architectures, LeNet [35] and ResNet-18 [19], as well as a distinct domain adaptation approach (Adversarial Discriminative Domain Adaptation, ADDA) [67]).

Results in Tab. 5 show that when training on only the source data, networks with *uniformity regularization* significantly outperform baseline models by as much as 18% on the target dataset. The gain in performance for ResNets and LeNets trained only on the source data demonstrates that such models disproportionately overfit to the training (or source) data, which we can alleviate via *uniformity regularization* to learn better data-agnostic features. Performance gains are also evident in ADDA, which operates under an adversarial training setting different from "Source Only" baseline models. In addition, ADDA with *uniformity regularization* achieves Zero-Shot Domain Adaptation performance close to that of a supervised learner trained directly on target data ("Target Only"), highlighting the strong benefits of *uniformity regularization* for transfer tasks especially under notable distribution shifts.

5.5. Uniform Priors benefit OOD Classification

In this section, we evaluate trained models on their ability to detect Out-of-Distribution (OOD) data. For that, we investigate two benchmarks. First, on CIFAR-10 data [32], we perform severe image augmentations using random translations of $[-4, 4]$ pixels, random rotations between $[-30, 30]$ degrees and scaling by a factor between $[0.75, 1.25]$, These transformations are physical transformations which preserve the semantics of the image. We then train the network on the standard dataset without any augmentations, and test three standard architectures [19, 75, 78] on the OOD testing set with the aforementioned augmentations. In addition to that, we also check OOD classification robustness on the corrupted CIFAR-100 variant, CIFAR-100-C, proposed in [20] to study network robustness.

Table 6. **Out-of-Distribution Classification**. Comparing OOD classification performance for various networks with mean accuracy and std over 5 seeds. OOD samples are generated via random translations, rotations, and scaling. The base task classification performance shows no performance degradation with $\mathcal{UR}$.

Task ↓	**ResNet-18**	+ $\mathcal{UR}$
OOD Accuracy	35.6 ± 1.2	**41.3** ± 1.3
Standard Test Accuracy	91.4 ± 0.6	91.8 ± 0.3
WideResNet-50	+ $\mathcal{UR}$	**ResNeXt-50**	+ $\mathcal{UR}$
39.6 ± 1.2	**43.9** ± 0.9	40.1 ± 0.8	**43.8** ± 1.1
94.7 ± 0.3	94.6 ± 0.3	95.3 ± 0.3	95.9 ± 0.5

Table 7. The mean error rate/Corruption Error (mCE) on the CIFAR-100-C benchmark (**lower is better**) indicates increased network robustness (lower error rates on corrupted data) with $\mathcal{UR}$.

		Noise			Blur			
Setting	Clean	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom
WRN-50	16.7	79.1	68.3	70.6	41.9	70.1	43.2	40.7
WRN-50 + $\mathcal{UR}$	**14.3**	**73.4**	**61.5**	**70.2**	**38.6**	**66.9**	**41.7**	**38.7**

Weather				Digital				
Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	mCE
49.6	51.8	18.4	20.1	36.4	39.9	50.6	50.9	46.8
43.8	**44.6**	**17.0**	**17.4**	**29.8**	**36.9**	**39.8**	**39.9**	**41.0**

Results are provided in Tables 6 and Table 7, respectively. In both cases, we find improvements in OOD classification accuracy. More specifically, our results in Tab. 6[3] highlight that across backbone networks, the ability to classify even under severe, unseen physical augmentations at test time is improved (while even retaining or partly improving base classification accuracy, see "*Standard Test Acc.*"), while Tab. 7 showcases improved robustness towards common image corruptions (such as noise and blurring) on CIFAR-100-C. Both results again detail the the disproportionate usefulness of *uniformity regularization* for transfer under different distribution shifts at test time.

[3]We note that the *relatively* high variance in Table 6 is due to the stochastic nature of the data augmentation techniques.

Table 8. **Single Scene View Synthesis** using NeRF [42] trained on the Co3D dataset [48].

Method ↓	PSNR	LPIPS	l_1 depth
NeRF	23.7	0.18	**0.37**
$+\mathcal{UR}$	**24.1**	**0.16**	**0.37**

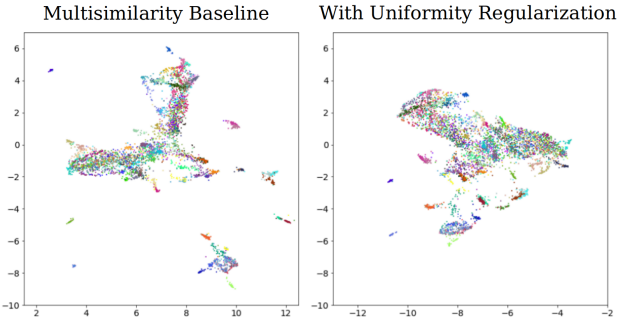


Figure 2. **Qualitative feature space study.** Evaluation of feature space changes caused by imposing a uniformity prior and producing a 2-d feature map using UMAP [40] in zero-shot Deep Metric Learning, showing representations spaces for a DML baseline model (Multisimilarity Loss [71]) with and without $\mathcal{UR}$. As the density increases with $\mathcal{UR}$, the generalization [51] improves (as shown quantitatively in section 5.7).

5.6. Uniform Priors and Neural Fields

Neural Radiance Fields (NeRFs) are trained to overfit to a scene while being able to generalize and produce novel views [42]. The need to generalize to novel views can be a bottleneck while training NeRFs, as they have shown to produce better results on known views in the training set, compared to unknown views that they are evaluated on [48]. Similar to before, we simply add the uniformity prior to the penultimate layer of the colour and densities MLPs used to train NeRFs. We then train a baseline NeRF both with and without uniformity regularization on the Co3D dataset [48], a recently proposed, large-scale dataset of "common objects" from multiple views. We report the results in Table 8, where we find, similar to before, that we are able to improve the NeRF baseline on two key metrics evaluated on novel views, namely PSNR (Peak-Signal-to-Noise-Ratio, [42]) and LPIPS (Learned Perceptual Image Patch Similarity, [80]) on the test set, showcasing the applicability of uniformity regularization to improve generalisability even for implicit generative modeling tasks.

5.7. How Uniform Priors change the feature space

Finally, we examine both qualitatively and quantitatively the influence of *uniformity regularization* on the feature space. For that, we look at the feature space changes in the Deep Metric Learning problem. Deep Metric Learning naturally lends itself to this study, as changes in the feature space are more strongly reflected in the underlying objective as well as the downstream performance, which primarily evaluate the goal of learning generalizing metric embedding spaces operating on top of learned image features.

Figure 2 qualitatively shows increased feature space density (less overclustering) when applying *uniformity regularization* and mapping to 2-d with UMAP [40], especially on the test data, showcasing that *uniformity regularization* has an impact on the feature distribution. As important and subtle changes are likely lost in the dimensionality reduction process, we also perform a quantitative evaluation of the actual feature space density, following the definition in [51]. Here, we find a 30% increase on the training feature density when applying *uniformity regularization*, which is consistent with [51] that link increased embedding space density on training data to improved test generalization. We note that [51] performed their embedding space studies on final embeddings produced by a linear mapping from the feature space, and thus believe insights to be transferable.

6. Conclusion

In this paper, we propose a regularization technique for generalization to novel tasks in Deep Learning. We present a simple and general solution, *uniformity regularization*, to reduce training bias and encourage networks to learn more reusable features. In a large experimental study, we show benefits across multiple, distinct domains studying varying degrees of fast adaptation and generalization such as Meta-Learning over both vision and language modalities, Deep Metric Learning, Zero-Shot Domain Adaptation and Out-of-Distribution Classification, and highlight the role of uniformity of the prior over learned features for generalization and adaptation. We further show that *uniformity regularization* achieves competitive performance on large-scale Few-Shot and Metric Learning tasks.

Broader Impact and Limitations We study the notion of uniform feature distributions for deep network training, offering a general regularization methods for Deep Learning tasks that require improved generalization. With this comes the chance for misuse in the respective domains. However, the improvements gained, while notable, are not significant enough to alter societal use in the respective areas. One of the limitations and future works for this paper is related to theoretical contributions. In the paper, we propose a practical framework for uniformity regularization. However, while we provide intuition and motivation and strong experimental support, we do not offer direct theoretical proofs for the experimental benefits we find.

Acknowledgements. Karsten Roth thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. 2

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1

[4] Xuefei Cao, Bor-Chun Chen, and Ser-Nam Lim. Unsupervised deep metric learning via auxiliary rotation loss. *CoRR*, abs/1911.07072, 2019. 2

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 2

[6] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *CoRR*, abs/2003.04390, 2020. 1, 2

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. pages 539–546. IEEE, 2005. 3

[8] Tristan Deleu, Tobias Würfl, Mandana Samiei, Joseph Paul Cohen, and Yoshua Bengio. Torchmeta: A Meta-Learning library for PyTorch, 2019. Available at: https://github.com/tristandeleu/pytorch-meta. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 6

[10] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss for deep metric learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, volume 12352 of *Lecture Notes in Computer Science*, pages 277–294. Springer, 2020. 6

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 1, 2, 4, 5

[12] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6249–6254. Association for Computational Linguistics, 2019. 6

[13] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018. 2

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[15] Ian Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In *ICLR2014*, 2014. 3, 7

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 1, 2, 6

[17] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 6

[18] Ben Harwood, BG Kumar, Gustavo Carneiro, Ian Reid, Tom Drummond, et al. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2829, 2017. 2

[19] 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*, pages 770–778, 2016. 1, 4, 6, 7

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 7

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 2

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 2

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 4

[24] Allan Jabri, Kyle Hsu, Abhishek Gupta, Ben Eysenbach, Sergey Levine, and Chelsea Finn. Unsupervised curricula for visual meta-reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 10519–10530, 2019. 2

[25] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11719–11727. Computer Vision Foundation / IEEE, 2019. 2

[26] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8166–8175. Computer Vision Foundation / IEEE, 2021. 2

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 4

[28] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[29] Louis Kirsch, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Improving generalization in meta reinforcement learning using learned objectives. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2

[30] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2452–2460, 2015. 1, 2

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2, 6

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 5, 7

[33] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019. 5

[34] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 3, 5, 7

[35] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 7

[36] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 2

[37] Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 4, 5, 6

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[39] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 2, 3

[40] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. 2020. 8

[41] Timo Milbich, Karsten Roth, Homanga Bharadhwaj, Samarth Sinha, Yoshua Bengio, Björn Ommer, and Joseph Paul Cohen. Diva: Diverse visual feature aggregation fordeep metric learning. *arXiv preprint arXiv:2004.13458*, 2020. 2

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 8

[43] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. *arXiv preprint arXiv:1712.09926*, 2017. 2

[44] Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[45] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. 2019. 6, 7

[46] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124, 2019. 2

[47] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 2

[48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 8

[49] Karsten Roth, Biagio Brattoli, and Bjorn Ommer. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8000–8009, 2019. 6

[50] Karsten Roth, Timo Milbich, and Björn Ommer. Pads: Policy-adapted sampling for visual similarity learning. *arXiv preprint arXiv:2003.11113*, 2020. 2, 6

[51] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. *arXiv preprint arXiv:2002.08473*, 2020. 1, 2, 5, 6, 8

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4

[53] Artsiom Sanakoyeu, Vadim Tschernezki, Uta Buchler, and Bjorn Ommer. Divide and conquer the embedding space for metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[54] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 2

[55] Alexander K Seewald. Digits-a dataset for handwritten digit recognition. *Austrian Research Institut for Artificial Intelligence Technical Report, Vienna (Austria)*, 2005. 7

[56] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5972–5981, 2019. 2

[57] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1, 2, 4, 5

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[59] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2

[60] Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms and experiments, 2018. 3

[61] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[62] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020. 1

[63] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning, 2020. 1

[64] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 5

[65] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017. 2

[66] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019. 2, 5, 6

[67] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 1, 2, 7

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 6

[69] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *Advances in neural information processing systems*, 2016. 2, 4, 5

[70] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020. 1, 2, 5

[71] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 2, 6, 8

[72] Yaming Wang, Jonghyun Choi, Vlad Morariu, and Larry S Davis. Mining discriminative triplets of patches for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1172, 2016. 2

[73] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2, 3, 6

[74] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 2, 6

[75] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4, 7

[76] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4, 2006. 3

[77] Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019. 2

[78] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 4, 7

[79] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018. 2, 6

[80] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[81] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019. 2