A Self-supervised GAN for Unsupervised Few-shot Object Recognition

Khoi Nguyen and Sinisa Todorovic
Oregon State University
Corvallis, OR 97330, USA
{nguyenkh,sinisa}@oregonstate.edu

***Abstract*—This paper addresses unsupervised few-shot object recognition, where all training images are unlabeled, and test images are divided into queries and a few labeled support images per object class of interest. The training and test images do not share object classes. We extend the vanilla GAN with two loss functions, both aimed at self-supervised learning. The first is a reconstruction loss that enforces the discriminator to reconstruct the probabilistically sampled latent code which has been used for generating the "fake" image. The second is a triplet loss that enforces the discriminator to output image encodings that are closer for more similar images. Evaluation, comparisons, and detailed ablation studies are done in the context of few-shot classification. Our approach significantly outperforms the state of the art on the Mini-Imagenet and Tiered-Imagenet datasets.**

I. Introduction

This paper presents a new deep architecture for unsupervised few-shot object recognition. In training, we are given a set of unlabeled images. In testing, we are given a small number K of *support images* with labels sampled from N object classes that do not appear in the training set (also referred to as unseen classes). Our goal in testing is to predict the label of a *query image* as one of these N previously unseen classes. A common approach to this N-way K-shot recognition problem is to take the label of the closest support to the query. Thus, our key challenge is to learn a deep image representation on unlabeled data such that it would in testing generalize well to unseen classes, so as to enable accurate distance estimation between the query and support images.

This problem is important as it appears in a wide range of applications. For example, we expect that leveraging unlabeled data could help few-shot image classification in domains with very few labeled images per class (e.g., medical images). Another example is online tracking of a previously unseen object in a video, initialized with a single bounding box. Unsupervised few-shot object recognition is different from the standard few-shot learning [1], [2] that has access to a significantly larger set of labeled images, allowing for episodic training [3]. Episodic training cannot be used in our setting with a few annotations.

There is scant work on unsupervised few-shot classification. Recent work [4], [5], [6] first identifies pseudo labels of unlabeled training images, and then uses the standard episodic training [3] on these pseudo labels. However, performance of these methods is significantly below that of counterpart approaches to supervised few-shot learning.

Motivated by the success of Generative Adversarial Networks (GANs) [7], [8], [9] to generalize well to new domains, we adopt and extend this framework with two new strategies for self-supervision [10]. A GAN is appropriate for our problem since it is a generative model aimed at learning the underlying image prior in an unsupervised manner, rather than discriminative image features which would later be difficult to "transfer" to new domains. As shown in Fig. 1, a GAN consists of a generator and discriminator that are adversarially trained such that the discriminator distinguishes between "real" and "fake" images, where the latter are produced by the generator from randomly sampled latent codes. We extend this framework by allowing the discriminator not only to predict the "real" or "fake" origins of the input image but also to output a deep image feature, which we will use later for unsupervised few-shot classification task. This allows us to augment the standard adversarial learning of the extended GAN with additional self-supervised learning via two loss functions – reconstruction loss and distance-metric triplet loss.

Our first contribution: By minimizing a reconstruction loss between the randomly sampled code and the discriminator's encoding of the "fake" image, we enforce the discriminator to explicitly capture the most relevant characteristics of the random codes that have been used to generate the corresponding "fake" images. In this way, the discriminator seeks to extract relevant features from images which happen to be "fake" but are guaranteed by the adversarial learning to be good enough to fool the discriminator of their origin. Thus, we use the randomly sampled codes not only for the adversarial learning but also as a "free" ground-truth for self-supervised learning [10]. From our experiments, the added reconstruction loss gives a significant performance improvement over the vanilla GAN.

Our second contribution: As shown in Fig. 1, we specify another type of self-supervised learning for our extended GAN so image encodings at the discriminator's output respect similarity of input images. While in general this additional self-supervision along with adversarial learning is expected to produce a better GAN model, it is particularly suitable for our subsequent distance-based image retrieval and few-shot image classification. In the lack of labeled data, for distance-metric learning, we resort to data augmentation. We take "real" training images and mask them with a patch whose placement controls similarity between the masked and original image,

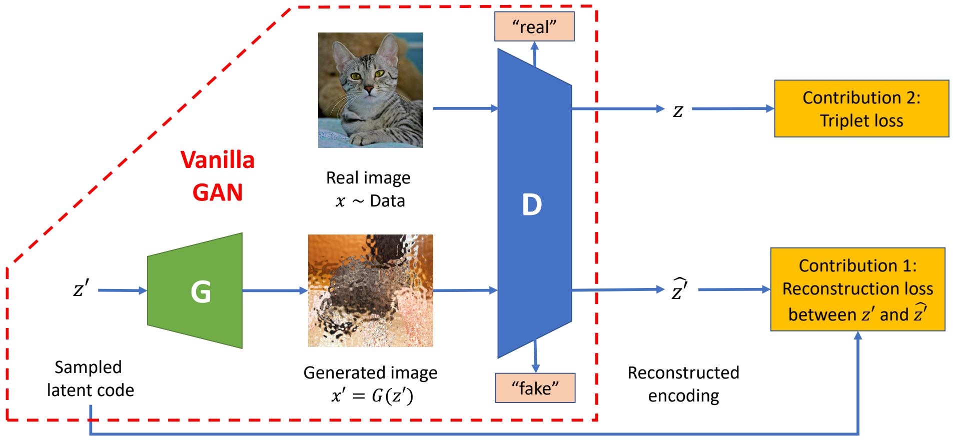

Fig. 1. We extend the vanilla GAN to learn an image encoding z on unlabeled data that will be suitable for subsequent few-shot image classification and image retrieval in new domains with very few annotations. Our extension integrates self-supervised and adversarial learning by the means of: (a) Reconstruction loss so the encoding $\widehat{z'}$ of a "fake" image is similar to the corresponding randomly sampled code z'; and (b) Deep metric learning so the image encodings z are closer for similar "real" images than for dissimilar ones.

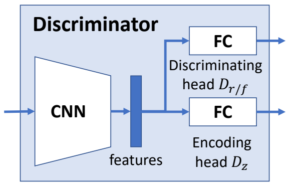

Fig. 2. Details of the discriminator from Fig. 1 with discriminating head and encoding head.

as shown in Fig. 3. Following the long track of research on object center-bias [11], [12], [13], [14] arguing that human attention usually focuses on an object in the image center, we declare the masked images more similar to the original if their masking patches fall in the image corners rather than on the image center, as the latter case is more likely to partially occlude the target object.

We use the above masking procedure to compile a set of image triplets that encode similarity relationships between the images. In a triplet, the anchor is a "real" image selected from the training dataset, the positive is a masked version of the anchor with the masking patch in one of the image corners, and the negative is another masked anchor but with the masking patch falling close to the image center. Thus, by construction, we constrain the triplet's similarity relationships relative to the anchor, as our training images are unlabeled. Given the set of triplets, we estimate the standard triplet loss for our GAN training. As our results show, the added distance-metric learning further improves the subsequent few-shot classification.

Our two contributions lead to significant performance gains in unsupervised few-shot image classification relative to recent unsupervised approaches on the Mini-Imagenet [3], [15] and Tiered-Imagenet [16] datasets.

In the rest of this paper: Sec. II reviews prior work, Sec. III specifies our approach, and Sec. IV presents our implementation details and experimental results.

II. Related Work

This section reviews closely related work.

Methods for **deep unsupervised learning** can be broadly divided into: deep clustering, self-supervised learning and generative models. Deep clustering iteratively trains a CNN [17], [18], [19]. In each iteration, deep feature extracted by the current CNN are clustered to produce pseudo-labels of training images. The pseudo-labels are then used for the standard supervised learning of the next CNN. Self-supervised methods seek to solve an auxiliary task, for which ground-truth can be easily specified [20], [21], [22], [23], [24]. Generative models seek to generate new images that look similar to images sampled from the underlying distribution. Recent generative models include (variational) auto-encoders [25], [26], and GANs [7], [8]. Our extended GAN belongs to the group of generative models enhanced with self-supervised learning.

For **unsupervised few-shot classification**, recent work uses unlabeled training data and very few labeled test data that do not share the same classes. These methods first assign pseudo labels to the training data by either image clustering [4] or treating each training example as having a unique class [5], [6]. Then, they apply the standard fully-supervised few-shot learning (e.g., [2], [1]) on the pseudo-labeled training data. We differ from these approaches in two ways. First, we do not estimate pseudo-labels, and do not use the common episodic training for fully-supervised few-shot learning, but seek to directly learn the underlying image prior distribution by integrating adversarial and self-supervised learning. Second, we ensure that our image representation respects similarity relationships of images.

Our problem is related to semi-supervised few-shot learning [16], [27]. These approaches first augment their labeled training set with unlabeled images, then, apply label propagation for knowledge transfer from the labeled to unlabeled set, and finally conduct fully-supervised few-shot learning on all training data. We cannot use this framework, as our labeled images have different classes from unlabeled ones, and even if they shared the same classes label, propagation from just one labeled example per class would be very difficult. Adversarial learning and self-supervised learning has been integrated for fully supervised few-shot learning [28], [29], but cannot be easily extended to our unsupervised setting.

Compared with very closely-related work [30], [31], [32], our approach significantly differs. [33], [31] use the same GAN but with a rotation-based self-supervision loss which we do not use. [32] uses a completely different triplet loss from ours. To form a triplet, we use the anchor, positive and negative from the very **same image**, whereas [32] uses the anchor and positive from the same image and the negative from a different image.

III. Our Approach

This section first specifies our extended GAN and its adversarial learning, Sec. III-A defines the reconstruction loss, Sec. III-B presents our masking procedure and triplet loss, and Sec. III-C describes our unsupervised training.

As shown in Figures 1 and 2, our extended GAN consists of the standard generator network G and a discriminator network D which is equipped with image encoding head D_z and the standard discriminating head $D_{\text{r/f}}$. For adversarial training, we probabilistically sample latent codes z' from a prior distribution, $z' \sim p(z')$, and generate corresponding "fake" images $x' = G(z')$. Both "real" training images x and "fake" images x' are passed to the discriminator for $D_{\text{r/f}}$ to predict their "real" or "fake" nature, $D_{\text{r/f}}(x) \in \mathbb{R}$.

For training D and G, we use the standard adversarial loss, specified as

$$\begin{aligned} L_D^{\text{adv}} &= \mathop{\text{E}}_{x\sim p_{\text{data}}(x)}[\max(0, 1 - D_{\text{r/f}}(x))] \\ &\quad + \mathop{\text{E}}_{z'\sim p(z')}[\max(0, 1 + D_{\text{r/f}}(G(z')))], \qquad (1) \\ L_G^{\text{adv}} &= - \mathop{\text{E}}_{z'\sim p(z')}[D_{\text{r/f}}(G(z'))], \qquad (2) \end{aligned}$$

where E denotes the expected value and $p_{\text{data}}(x)$ is a prior distribution of unlabeled "real" training images x. As shown in [34], optimizing (1) and (2) is equivalent to minimizing the reverse KL divergence.

For sampling latent codes, $z' \sim p(z')$, as in related work [7], [8], [9], we assume that all elements of z' are i.i.d.. For the d-dimensional latent code z', we have studied several definitions of the prior $p(z')$, including the continuous uniform distribution $p(z') = U[-1, 1]^d$, the corresponding discrete Bernoulli distribution with the equal probability of 0.5 for the binary outcomes "1" and "-1" of elements of z', and the Gaussian distribution with the mean zero and variance 1, $p(z') = \mathcal{N}(0, 1)^d$. All these definitions of $p(z')$ give similar performance in our experiments. Sampling from the uniform distribution is justified, because many object classes appearing in training images are likely to share the same latent features, so the presence of these features encoded in z and z' is likely to follow the uniform distribution.

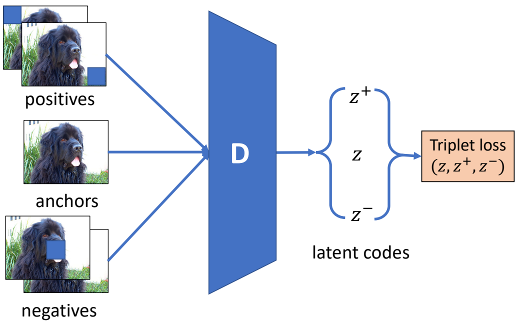

Fig. 3. Self-supervised learning of our extended GAN by minimizing the triplet loss of ⟨anchor, positive, negative⟩ images, where the positives and negatives are generated via appropriately masking the anchor. D is the discriminator show in Fig. 2

A. Reconstruction Loss

We extend the above specified adversarial learning of our GAN by enforcing the discriminator's encoding head D_z to reconstruct randomly sampled latent codes $z' \sim p(z')$ from the corresponding "fake" images, $\hat{z}' = D_z(G(z')) \in \mathbb{R}^d$, as illustrated in Fig. 1. Thus, we use z' as a "free" label for the additional self-supervised training of D and G along with the adversarial learning. Minimizing a reconstruction loss between z' and $\hat{z}'$ enforces D_z to be trained on "free" labels of the corresponding "fake" images – the knowledge that will be later transferred for encoding "real" images by D_z when "fake" images become good enough to "fool" $D_{\text{r/f}}$.

A difference between z' and its reconstruction $\hat{z}' = D_z(G(z'))$ is penalized with the mean-squared error (MSE) reconstruction loss:

$$L_D^{\text{mse}} = L_G^{\text{mse}} = \|\hat{z}' - z'\|_2^2, \qquad (3)$$

when elements of z' are sampled from the uniform or Gaussian distribution, $p(z') = U[-1, 1]^d$ or $p(z') = \mathcal{N}(0, 1)^d$, or with the binary cross-entropy (BCE) reconstruction loss:

$$\begin{aligned} L_D^{\text{bce}} = L_G^{\text{bce}} = -\frac{1}{d} \sum_{m=1}^{d} [\frac{1 + z'_m}{2} \cdot \log \sigma(\hat{z}'_m) \\ + (1 - \frac{1 + z'_m}{2}) \cdot \log(1 - \sigma(\hat{z}'_m))], \qquad (4) \end{aligned}$$

when elements of z', $z'_m \in \{-1, 1\}$, are sampled from the discrete Bernoulli distribution, $p(z'_m = \pm 1) = 0.5$; $\sigma(\cdot)$ denotes the sigmoid function.

B. Image Masking and Triplet Loss

We additionally train the discriminator to output image representations that respect image similarity relationships, such that $z_i = D_z(x_i)$ and $z_j = D_z(x_j)$ are closer than z_i and $z_k = D_z(x_k)$ when images x_i and x_j are more similar than x_i and x_k.

As ground-truth similarity relationships are not provided for our training images, we resort to data augmentation using the following masking procedure. Each training image generates a set of its masked versions by exhaustively placing a masking patch on a regular grid covering the image. The masking patch brightness is uniform and equal to the pixel average of the original image. We have experimented with various patch sizes and shapes, and various grids. A good trade-off between complexity and performance for 64×64 training images that we use is obtained for the 16×16 square patch and 4×4 regular grid.

Given the training dataset and its masked images, we compile a set of image triplets $\langle$anchor, positive, negative$\rangle$. The anchor $z = D_z(x)$ is an image from the training set. The positive is one of four masked versions of the anchor image $\{z_i^+ = D_z(x_i^+) : i = 1, \dots, 4\}$ whose masking patch falls in the image corner – namely, the top-left, top-right, bottom-left, and bottom-right corner. The negative is one of the remaining masked versions of the anchor image $\{z_j^- = D_z(x_j^-) : j = 1, 2, \dots\}$ whose masking patch falls on central locations on the grid. In the positives, the relatively small masking patch masks very little to no foreground of the anchor image. On the other hand, in the negatives, the masking patch is very likely to partially occlude foreground of the anchor image. Therefore, by construction, we constrain image similarity relationships to pertain to the image's foreground, such that the positives should be closer to the anchor than the negatives.

We use the set of triplets $\{\langle z, z_i^+, z_j^- \rangle\}$ to estimate the following triplet loss for the additional distance-metric learning of our GAN:

$$L_D^{\text{triplet}} = \max[0, \max_i \Delta(z, z_i^+) - \min_j \Delta(z, z_j^-) + \rho]. \quad (5)$$

where $\rho \geq 0$ is a distance margin, and Δ is a distance function. In this paper, we use the common cosine distance:

$$\Delta(z, z') = 1 - \frac{z^\top z'}{\|z\|_2 \cdot \|z'\|_2}. \quad (6)$$

Algorithm 1: Our Unsupervised Training

```
Input = $\{T_1, T_2, \mathcal{T}, \beta, \gamma, \lambda\}$= numbers of training iterations, and 3
non-negative hyper parameters.
                                                        ▷ First stage (1)
for $t_1 = 1, \dots, T_1$ do
    Sample $z' \sim p(z')$ and reconstruct $\hat{z'} = D_z(G(z'))$;
    Compute: $L_G^{\text{adv}}$ as in (2), and $L_G^{\text{mse}}$ as in (3) or $L_G^{\text{bce}}$ as in (4);
    Update $G$ by back-propagating the total loss:
        $L_G^{(1)} = L_G^{\text{adv}} + \beta L_G^{\text{bce}}$ or $L_G^{(1)} = L_G^{\text{adv}} + \beta L_G^{\text{mse}}$.
    for $\tau = 1, \dots, \mathcal{T}$ do
        Sample $z' \sim p(z')$ and reconstruct $\hat{z'} = D_z(G(z'))$;
        Sample a real image $x \sim p_{\text{data}}(x)$;
        Compute: $L_D^{\text{adv}}$ as in (1);
        Compute: $L_G^{\text{mse}}$ as in (3) or $L_D^{\text{bce}}$ as in (4);
        Update $D^{(1)}$ by back-propagating the total loss:
            $L_D^{(1)} = L_D^{\text{adv}} + \gamma L_D^{\text{bce}}$ or $L_D^{(1)} = L_D^{\text{adv}} + \gamma L_D^{\text{mse}}$.
    end for
end for
                                                        ▷ Second stage (2)
for $t_2 = 1, \dots, T_2$ do
    Sample a real image $x \sim p_{\text{data}}(x)$ and take it as the anchor;
    Generate the masked positives and negatives of the anchor;
    Form the set of triplets for the anchor;
    Compute $L_D^{\text{triplet}}$ as in (5);
    Update $D^{(2)}$ by back-propagating the total loss:
        $L_D^{(2)} = L_D^{\text{triplet}} + \lambda \|D_z^{(1)}(x) - D_z^{(2)}(x)\|_2^2$.
end for
Take $D^{(2)}$ as the learned discriminator $D$.
```

C. Our Unsupervised Training

Alg. 1 summarizes our unsupervised training that integrates adversarial learning with distance-metric learning and latent-code reconstruction. For easier training of D and G, we divide learning in two stages. First, we perform adversarial learning constrained with the latent-code reconstruction regularization over $t_1 = 1, \dots, T_1$ iterations. In every iteration, G is optimized once and D is optimized multiple times over $\tau = 1, \dots, \mathcal{T}$ iterations ($\mathcal{T} = 3$). After convergence of the first stage ($T_1 = 50,000$), the resulting discriminator is saved as $D^{(1)}$. In the second training stage, we continue with distance-metric learning in $t_2 = 1, \dots, T_2$ iterations ($T_2 = 20,000$), while simultaneously regularizing that the discriminator updates do not significantly deviate from the previously learned $D^{(1)}$.

IV. Experiments

Datasets: For evaluation on unsupervised few-shot classification, we follow [4], [5], [6], and evaluate on two benchmark datasets: Mini-Imagenet [3] and Tiered-Imagenet [16]. Our training is unsupervised, starts from scratch, and does not use other datasets for pre-training.

Mini-Imagenet consists of 100 randomly chosen classes from ILSVRC-2012 [35]. As in [8], [5], [4], these classes are randomly split into 64, 16, and 20 classes for training, validation, and testing, respectively. Each class has 600 images of size 84×84. Tiered-Imagenet consists of 608 classes of 84×84 images from ILSVRC-2012 [35], grouped into 34 high-level categories. These are divided into 20, 6 and 8 categories for meta-training, meta-validation, and meta-testing. This corresponds to 351, 97 and 160 classes for meta-training, meta-validation, and meta-testing respectively. Tiered-Imagenet minimizes the semantic similarity between the splits compared to Mini-Imagenet.

Evaluation metrics: For few-shot classification, we first randomly sample N classes from the test classes and K examples for each sampled class, and then classify query images into these N classes. We report the average accuracy over 1000 episodes with 95% confidence intervals of the N-way K-shot classification.

Specifically, we are given support images x_s with labels $y_s \in L_{\text{test}}$ sampled from $N = |L_{\text{test}}|$ classes which have not been seen in training. Each class has K examples, $K \leq 5$. Given a query image, x_q, we predict a label $\hat{y}_q \in L_{\text{test}}$ of the query as follows. After computing deep representations $z_q = D_z(x_q)$ and $z_s = D_z(x_s)$ of the query and support images, for every class $n = 1, \dots, N$, we estimate its mean prototype vector c_n as $c_n = \frac{1}{K} \sum_{s, y_s = n} z_s$, and take the label of the closest c_n to z_q as our solution:

$$\hat{y}_q = \hat{n} = \arg\min_n \Delta(z_q, c_n), \quad (7)$$

where Δ is a distance function (e.g., (6)). The same formulation of few-shot recognition is used in [1].

Implementation details: Our implementation is in Pytorch [36]. Our backbone GAN is the Spectral Norm GAN (SN-

GAN) [9] combined with the self-modulated batch normalization [33]. All images are resized to 64×64 pixels, since the SN-GAN cannot reliably generate higher resolution images beyond 64×64. There are 4 blocks of layers in both G and D. The latent code z' and image representation z have length $d = 128$. We use an ADAM optimizer [37] with the learning rate of $5e^{-4}$. We empirically observe convergence of the first and second training stages after $T_1 = 50000$ and $T_2 = 10000$ iterations, respectively. D is updated in $\mathcal{T} = 3$ iterations for every update of G. In all experiments, we set $\gamma = 1, \beta = 1, \lambda = 0.2, \rho = 0.5$ as they are empirically found to give the best performance. In the first and the second training stages, the mini-batch size is 128 and 32, respectively. Our image masking places a 16×16 patch at 4×4 locations of the regular grid in the original image, where the patch brightness is equal to the average of image pixels. It is worth noting that we do not employ other popular data-augmentation techniques in training (e.g., image jittering, random crop, etc.).

Ablation study: The following variants test how individual components of our approach affect performance:

- T: An architecture that is not a GAN, but consists of only the discriminator network from the SN-GAN [9], and the discriminator's encoding head D_z is trained on the triplet loss only. This variant tests how distance-metric learning affects performance without adversarial learning.
- Gc and Gd: SN-GAN [9] with continuous uniformly distributed and discrete Bernoulli-distributed elements of the latent code z', respectively.
- GcM and GdB: Gc and Gd are extended with the MSE and BCE reconstruction loss, respectively.
- GcT1 and GcT2 = Gc + T: Gc is extended with the triplet loss, and the discriminator is trained in a single stage with total loss $L_D = L_D^{\text{adv}} + \gamma L_D^{\text{triplet}}$ and in two stages as specified in Alg. 1, respectively. These two variants compare performance of single-stage and two-stage training.
- GdT1 and GdT2 = Gd + T: Gd is extended similarly as Gc for GcT1 and GcT2.
- GcMT1 and GcMT2 = Gc + MSE + T: Gc is extended with the MSE reconstruction loss and the triplet loss, and the discriminator is trained in a single stage with total loss $L_D = L_D^{\text{adv}} + \gamma L_D^{\text{mse}} + \gamma L_D^{\text{triplet}}$ and in two stages as specified in Alg. 1, respectively.
- GdBT1 and GdBT2 = Gd + BCE + T: Gd is extended the BCE reconstruction loss and the triplet loss, and the discriminator is trained in a single stage with total loss $L_D = L_D^{\text{adv}} + \gamma L_D^{\text{bce}} + \gamma L_D^{\text{triplet}}$ and in two stages as specified in Alg. 1, respectively.

Table. I presents our ablation study on the tasks of unsupervised 1-shot and 5-shot image classification on Mini-Imagenet. As can be seen, the variants with discrete Bernoulli-distributed elements of z', on average, achieve better performance by 0.5% than their counterparts with continuously distributed latent codes z'. Incorporating the reconstruction loss improves performance up to 9% over the variants whose discriminator does not reconstruct the latent codes. While the variant T is the worst, integrating the triplet loss with adversarial learning outperforms the variants Gc and Gd which use only adversarial learning by more than 8%. Relative to Gc and Gd, larger performance gains are obtained by additionally using the reconstruction loss in GcM and GdB than using the triplet loss in GcT2 and GdT2. Finally, the proposed two-stage training in Alg. 1 gives better results than a single-stage training, due to, in part, difficulty to optimize hyper parameters. The bottom three variants in Table. I evaluate GdBT2 for different sizes of the masking patch. From Table. I, GdBT2 gives the best results when the masking patch has size 16×16 pixels, and we use this model for comparison with prior work.

Comparison with state of the art: Tab II compares our GdBT2 with the state of the art on the tasks of unsupervised 1-shot and 5-shot image classification on Mini-Imagenet and Tiered-Imagenet. For fair comparison, we follow the standard label assignment to query images as in [1]. As can be seen, we significantly outperform the state of the art [4] by 9% in 1-shot and nearly 13% in 5-shot settings of Mini-Imagenet dataset. At the bottom, Tab. II also shows results of recent fully-supervised approaches to few-shot classification, as their performance represents our upper bound. Surprisingly, our GdBT2 achieves significantly outperforms the fully supervised ProtoNets [1]. Our approach's performance can be even further boosted in practice since we can make use of abundant unsupervised data while supervised approaches are not applicable.

Qualitative Results: Fig. 4 illustrates our masking procedure for generating positive and negative images in the triplets for distance-metric learning. In each row, the images are organized from left to right by their estimated distance to the original (unmasked) image in the descending order, where the rightmost image is the closest. From Fig. 4, masked images

TABLE I
AVERAGE ACCURACY OF FEW-SHOT IMAGE CLASSIFICATION ON TEST CLASSES IN MINI-IMAGENET. GDBT2 WITH THE 16×16 MASKING PATCH GIVES THE BEST RESULTS.

Ablations	1-shot	5-shot
T	31.23 ± 0.46	41.91 ± 0.53
Gc	34.52 ± 0.57	44.24 ± 0.72
Gd	34.84 ± 0.68	44.73 ± 0.67
GcM	43.49 ± 0.76	57.62 ± 0.73
GdB	43.51 ± 0.77	57.94 ± 0.76
GcT1	41.95 ± 0.47	50.62 ± 0.54
GdT1	42.13 ± 0.52	51.39 ± 0.63
GcMT1	45.13 ± 0.63	58.87 ± 0.69
GcBT1	45.43 ± 0.78	58.96 ± 0.72
GcT2	42.36 ± 0.45	52.76 ± 0.52
GdT2	43.13 ± 0.53	53.39 ± 0.78
GcMT2	46.72 ± 0.73	60.92 ± 0.74
GdBT2	**48.28 ± 0.77**	**66.06 ± 0.70**
GdBT2 8×8	46.23 ± 0.77	60.46 ± 0.64
GdBT2 16 × 16	**48.28 ± 0.77**	**66.06 ± 0.70**
GdBT2 32×32	44.40 ± 0.64	57.93 ± 0.49

TABLE II
UNSUPERVISED FEW-SHOT CLASSIFICATION ON MINI-IMAGENET AND TIERED-IMAGENET. AT THE BOTTOM, WE SHOW RESULTS OF RECENT FULLY-SUPERVISED APPROACHES TO FEW-SHOT CLASSIFICATION AS OUR UPPER BOUND.

	Mini-Imagenet, 5-way		**Tiered-Imagenet, 5-way**	
Unsupervised Methods	1-shot	5-shot	1-shot	5-shot
SN-GAN [9]	34.84 ± 0.68	44.73 ± 0.67	35.57 ± 0.69	49.16 ± 0.70
AutoEncoder [26]	28.69 ± 0.38	34.73 ± 0.63	29.57 ± 0.52	38.23 ± 0.72
Rotation [38]	35.54 ± 0.47	45.93 ± 0.62	36.90 ± 0.54	51.23 ± 0.72
BiGAN kNN [8]	25.56 ± 1.08	31.10 ± 0.63	-	-
AAL-ProtoNets [5]	37.67 ± 0.39	40.29 ± 0.68	-	-
UMTRA + AutoAugment [6]	39.93	50.73	-	-
CACTUs-ProtoNets [4]	39.18 ± 0.71	53.36 ± 0.70	-	-
Our GdBT2	**48.28 ± 0.77**	**66.06 ± 0.70**	**47.86 ± 0.79**	**67.70 ± 0.75**
Fully-supervised Methods				
ProtoNets [1]	46.56 ± 0.76	62.29 ± 0.71	46.52 ± 0.72	66.15 ± 0.74

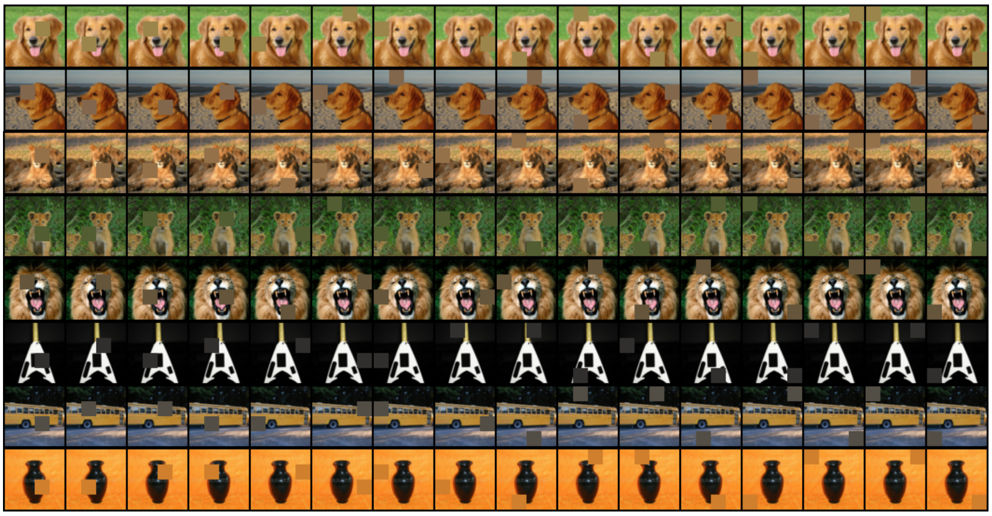

Fig. 4. Our image masking with rectangular patches for Mini-Imagenet. In every row, the images are organized from left to right in the descending order by their estimated distance to the original (unmasked) image.

that are the closest to the original have the masking patch in the image corner, and thus are good candidates for the positives in the triplets. Also, when the masking patch covers central image areas the resulting masked images have greater distances from the original, and are good candidates for the negatives in the triplets.

V. Conclusion

We have addressed unsupervised few-shot object recognition, where all training images are unlabeled and do not share classes with test images. We have extended the vanilla GAN so as to integrate the standard adversarial learning with our two new strategies for self-supervised learning. The latter is specified by enforcing the GAN's discriminator to: (a) reconstruct the randomly sampled latent codes, and (b) produce image encodings that respect similarity relationships of images. Results of an extensive ablation study on few-shot classification demonstrate that integrating: (i) triplet loss with adversarial learning outperforms the vanilla GAN by more than 8%, (ii) reconstruction loss with adversarial learning gives a performance gain of more than 9%, and (iii) both triplet and reconstruction losses with adversarial learning improves performance by 13%. In unsupervised few-shot classification, we outperform the state of the art by 9% on Mini-Imagenet in 1-shot and 13% in 5-shot settings.

Acknowledgement

This work was supported in part by DARPA XAI Award N66001-17-2-4029 and DARPA MCS Award N66001-19-2-4035.

References

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087. 1, 2, 4, 5, 6

[2] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135. 1, 2

[3] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638. 1, 2, 4

[4] K. Hsu, S. Levine, and C. Finn, "Unsupervised learning via meta-learning," *arXiv preprint arXiv:1810.02334*, 2018. 1, 2, 4, 5, 6

[5] A. Antoniou and A. Storkey, "Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation," *arXiv preprint arXiv:1902.09884*, 2019. 1, 2, 4, 6

[6] S. Khodadadeh, L. Bölöni, and M. Shah, "Unsupervised meta-learning for few-shot image and video classification," *arXiv preprint arXiv:1811.11819*, 2018. 1, 2, 4, 6

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 1, 2, 3

[8] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016. 1, 2, 3, 4, 6

[9] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018. 1, 3, 5, 6

[10] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *arXiv:1902.06162*, 2019. 1

[11] A. Borji and J. Tanner, "Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1214–1226, 2015. 2

[12] J. M. Henderson, "Eye movement control during visual object processing: effects of initial fixation position and semantic constraint." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 47, no. 1, p. 79, 1993. 2

[13] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *Journal of vision*, vol. 10, no. 8, pp. 20–20, 2010. 2

[14] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of vision*, vol. 8, no. 3, pp. 3–3, 2008. 2

[15] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *International Conference on Learning Representation*, 2016. 2

[16] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018. 2, 4

[17] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149. 2

[18] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, 2016, pp. 478–487. 2

[19] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742. 2

[20] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430. 2

[21] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666. 2

[22] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84. 2

[23] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5898–5906. 2

[24] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1058–1067. 2

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010. 2, 6

[27] Q. Sun, X. Li, Y. Liu, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," *arXiv preprint arXiv:1906.00562*, 2019. 2

[28] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 2365–2374. 3

[29] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," *arXiv preprint arXiv:1906.05186*, 2019. 3

[30] T. Chen, X. Zhai, M. Ritter, M. Lucic, and N. Houlsby, "Self-supervised gans via auxiliary rotation loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 154–12 163. 3

[31] N.-T. Tran, V.-H. Tran, B.-N. Nguyen, L. Yang *et al.*, "Self-supervised gan: Analysis and improvement with multi-class minimax game," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 232–13 243. 3

[32] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060. 3

[33] T. Chen, M. Lucic, N. Houlsby, and S. Gelly, "On self modulation for generative adversarial networks," *arXiv preprint arXiv:1810.01365*, 2018. 3, 5

[34] J. H. Lim and J. C. Ye, "Geometric gan," *arXiv preprint arXiv:1705.02894*, 2017. 3

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015. 4

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *NIPS-W*, 2017. 4

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[38] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018. 6