Self-training solutions for the ICCV 2023 GeoNet Challenge

Team Name: CASIA-TIM (Lijun Sheng, Zhengbo Wang, and Jian Liang ✉)
Affiliation: Institute of Automation, Chinese Academy of Sciences (CASIA)
Contact author: Jian Liang (liangjian92@gmail.com)

Challenge Website Challenge Leaderboard

Abstract

GeoNet is a recently proposed domain adaptation benchmark consisting of three challenges (i.e., GeoUniDA, GeoImNet, and GeoPlaces). Each challenge contains images collected from the USA and Asia where there are huge geographical gaps. Our solution adopts a two-stage source-free domain adaptation framework with a Swin Transformer backbone to achieve knowledge transfer from the USA (source) domain to Asia (target) domain. In the first stage, we train a source model using labeled source data with a re-sampling strategy and two types of cross-entropy loss. In the second stage, we generate pseudo labels for unlabeled target data to fine-tune the model. Our method achieves an H-score of 74.56% and ultimately ranks 1st in the GeoUniDA challenge. In GeoImNet and GeoPlaces challenges, our solution also reaches a top-3 accuracy of 64.46% and 51.23% respectively. The key aspects are shown below.

1. *Team name, primary contact/author, and list of all other participants.*

   CASIA-TIM (Lijun Sheng, Zhengbo Wang, and Jian Liang) (Contact author: liangjian92@gmail.com)

2. *Key highlights/salient aspects of your approach.*

   source-free domain adaptation, self-training strategy.

3. *Total size and kinds of data used in both pre-training, fine-tuning and training phases.*

   We use ImageNet-1k pre-trained model. The USA train split is used in source training and the Asia train split w/o labels is used in target adaptation.

4. *Total model sizes (# of parameters, type of architecture) and training strategies.*

   Swin-B (88M parameters). Fine-tuning on source domain & self-training on unlabeled target domain.

5. *Complete list of foundational models, if any, used in training your algorithm.*

   N/A. Only ImageNet-1k pre-trained Swin-B is used.

1. Background

Source-free domain adaptation (SFDA) [2] is a popular paradigm to achieve domain adaptation with only access to pre-trained source models and unlabeled target data. Due to its privacy protection of source data and competitive performance compared to conventional domain adaptation, SFDA gains more attention in the transfer learning tasks with distribution shift. SFDA methods always contain a source training stage and a target adaptation stage. They fine-tune the pre-trained source model with only unlabeled target data through unsupervised learning strategies, such as pseudo-labeling, consistency, and clustering. Note that SFDA methods are not allowed to access the source dataset during adaptation.

Swin Transformer [3] is a strong backbone for a broad range of computer vision tasks. Its efficiency and flexibility benefit from the design of shifted windows and the hierarchical architecture. Especially, Swin Transformer base model (Swin-B) obtains an accuracy of 86.4% on ImageNet-1K benchmark. Due to its competitive performance on ImageNet-1K benchmark and attractive generalization ability, we choose Swin-B as our backbone network in all experiments.

2. Framework

2.1. Overview

We adopt a two-stage source-free domain adaptation framework [2] for all three challenges in GeoNet benchmark. Our solution consists of a source training stage and a target adaptation stage. In the source training stage, we pre-train a source model using labeled USA images with the initialization of ImageNet-1k pretrained Swin-B model. In the target adaptation stage, pseudo labels of unlabeled Asia images are introduced and the source model is fine-tuned through the self-training strategy to improve its performance in Asia domain. Due to the source training stages being the same for all challenges, we describe its process in the next subsection. Target adaptation details for the three challenges are provided in their respective sections.

2.2. Source Training

Inspired by SHOT [2], our model $f(x)$ consists of a feature extractor $g : \mathcal{X} \rightarrow \mathbb{R}^d$ and a classifier module $h : \mathbb{R}^d \rightarrow \mathbb{R}^K$, i.e., $f(x) = h(g(x))$. d is the dimension of feature space which is set to 256 and K is the number of categories. Feature extractor g consists of a Swin-B backbone, a linear layer for dimension transformation, and a batch normalization (BN) layer. The classifier module h is designed as a standard linear layer whose output dimension equals the category number.

In the source training stage, we are given source training dataset $\{x_s^i, y_s^i\}_{i=1}^{n_s}$ from USA domain $\mathcal{D}_s$ to train a source model $f_s(x) = h_s(g_s(x))$. We adopt cross-entropy loss with the label smoothing technique for better generalization ability. The objective function is given by,

$$\mathcal{L}_{src}^{ls} = -\mathbb{E}_{(x_s,y_s)} \sum\nolimits_{k=1}^{K} q_k^{ls} \log \delta_k(f_s(x_s)), \tag{1}$$

where $q_k^{ls} = (1-\alpha)q_k + \alpha/K$ is the smoothed label and α is set to 0.1, and δ denotes soft-max operation.

Also, we use the exponential moving average (EMA) mechanism to maintain an EMA model that has better intermediate representations. EMA model parameters are updated once per epoch using an EMA mechanism with a smoothing coefficient of 0.95. We use cross-entropy loss between the model prediction and the EMA model prediction, which is given as,

$$\mathcal{L}_{src}^{EMA} = -\mathbb{E}_{(x_s,y_s)} \sum\nolimits_{k=1}^{K} \delta_k(f_s^{EMA}(x_s)) \log \delta_k(f_s(x_s)). \tag{2}$$

Considering that the category distribution of the source domain may be inconsistent with the target, we introduce a re-sampling strategy to mitigate the negative impact of the potential inconsistency. For the c-th class, the weight $\mathcal{W}_c$ is calculated by the reciprocal of the proportion of the category sample number N_i to the total:

$$\mathcal{W}_c = \frac{\sum_{i=1}^{K} \mathcal{N}_i}{\mathcal{N}_c}. \tag{3}$$

To summarize, during the source training stage, a re-sampling strategy is introduced and the full objective is as,

$$\mathcal{L}_{src} = \mathcal{L}_{src}^{ls} + \lambda_{EMA}\mathcal{L}_{src}^{EMA}, \tag{4}$$

where λ_{EMA} stays at zero for the first 40% of iterations, and then switches to one for the rest of the training process.

3. Challenge track: UniDA on GeoUniDA

3.1. Task

GeoUniDA [1] is a universal domain adaptation challenge with a huge geographical gap. Both source and target domains have 62 shared (a.k.a., known) classes and 138 private (a.k.a., unknown) classes. The goal of the task is to classify the samples from the shared classes correctly and mark the samples from the target private classes as "unknown". The performance is evaluated by the H-score metric which is the harmonic mean of the known accuracy and the binary unknown accuracy.

3.2. Target Adaptation

After employing the source training process in Sec. 2.2, we fine-tune the source model with unlabeled target dataset $\{x_t^i\}_{i=1}^{n_t}$. Following SHOT [2], we use the source model as initialization, freeze the classifier module h_t, and optimize the feature extractor g_t.

We divide the target domain samples into known set $\mathcal{S}_k$ and unknown set $\mathcal{S}_u$ for deploying different optimization strategies. Two thresholds (i.e., τ_{high} and τ_{low}) are introduced to define the boundary of two sets. $\mathcal{S}_k$ consists of samples whose entropy of prediction is lower than τ_{low}. Samples with entropy greater than τ_{high} belong to $\mathcal{S}_u$:

$$\begin{aligned} \mathcal{S}_k &= \{x_t \| \mathcal{H}(x_t) \leq \tau_{low}\}, \\ \mathcal{S}_u &= \{x_t \| \mathcal{H}(x_t) \geq \tau_{high}\}, \end{aligned} \tag{5}$$

where $\mathcal{H}(x_t) = -\sum_{k=1}^{K} \delta_k(h_t(g_t(x_t))) \log \delta_k(h_t(g_t(x_t)))$ denotes the entropy of target data x_t.

For stable training, we update τ_{high} and τ_{low} from initial value 0.5 consecutively:

$$\begin{aligned} \tau_{high} &= 0.5 + 0.2 \times \zeta, \\ \tau_{low} &= 0.5 - 0.2 \times \zeta, \end{aligned} \tag{6}$$

where $\zeta \in (0, 1)$ is a variable that increases uniformly.

For samples in $\mathcal{S}_k$, we treat them as known samples and calculate cross-entropy loss between their predictions and pseudo labels $\hat{y}_t$:

$$\begin{aligned} \mathcal{L}_k(g_t) &= -\mathbb{E}_{(x_t,\hat{y}_t) \in \mathcal{S}_k \times \hat{\mathcal{Y}}_t} \log \delta_{\hat{y}}(h_t(g_t(x_t))), \\ \hat{y}_t &= \underset{k}{\text{argmax}} \ \delta_k(h_t(g_t(x_t))). \end{aligned} \tag{7}$$

For samples in $\mathcal{S}_u$, we treat them as unknown samples and try to maximize their entropy to improve the outlier recognition ability:

$$\mathcal{L}_u(g_t) = -\mathbb{E}_{x_t \in \mathcal{S}_u} \mathcal{H}(x_t). \tag{8}$$

To summarize, during the target adaptation stage, the full objective is as,

$$\mathcal{L}_{tar}^{UniDA}(g_t) = \mathcal{L}_k(g_t) + \alpha \mathcal{L}_u(g_t), \tag{9}$$

where α is a trade-off hyperparameter which is set to 0.3.

Table 1. H-score (%) on Asia test split and challenge test set of GeoUniDA challenge ('old/new' refers to last/best source model checkpoint, 'source/target' refers to performance before/after target adaptation stage).

Method	Asia test split	Challenge (eval.ai)
Baseline	-	53.59
old source	61.47	70.24
old adapt	64.34	74.43
new source	62.03	71.08
new adapt	64.45	**74.56**

3.3. Result

We evaluate our solution on Asia test split and challenge test set of GeoUniDA challenge and results are shown in Table 1. We use the last source checkpoint as the source model in 'old' methods and use the best one in 'new' methods. Besides, 'source' refers to the performance of the source model, and 'adapt' refers to the performance after target adaptation. Based on the provided results, our solution achieves an H-score of 74.56% on the challenge test set which is the 1st place in the leaderboard.

4. Challenge track: UDA on GeoPlaces

4.1. Task

GeoPlaces [1] is an unsupervised domain adaptation challenge with a large huge geographical gap. It consists of images from 204 places across two domains (i.e., USA and Asia). The goal of the task is Asian place recognition by utilizing the labeled set from the USA and the unlabeled data from Asia. The performance is evaluated by the top-3 accuracy.

4.2. Target Adaptation

In the target adaptation stage, the model is fine-tuned with unlabeled target dataset $\{x_t^i\}_{i=1}^{n_t}$. Following SHOT [2], we use the source model as initialization, freeze the classifier module h_t, and optimize the feature extractor g_t.

The self-training strategy calculates the cross-entropy loss between the model prediction and the pseudo label. Due to the difficulty of GeoPlaces challenge, the accuracy of the pseudo label is too low Thus We choose both the first and second highest categories as double pseudo labels to provide supervision signals.

The generation of two pseudo labels of $\{x_t^i\}_{i=1}^{n_t}$ is as,

$$\begin{aligned} \hat{y}_t^{1st} &= \underset{k}{\text{argmax}}\ \delta_k(f_s(x_t)), \\ \hat{y}_t^{2nd} &= \underset{k\neq\hat{y}_t^{1st}}{\text{argmax}}\ \delta_k(f_s(x_t)). \end{aligned} \tag{10}$$

Based on double pseudo labels $\hat{y}_t^{1st}, \hat{y}_t^{2nd}$, the cross-entropy loss is calculated as,

$$\begin{aligned} \mathcal{L}_{tar}^{Places}(g_t) = &-\beta\, \mathbb{E}_{(x_t,\hat{y}_t^{1st})\in\mathcal{X}_t\times\hat{\mathcal{Y}}_t} \log \delta_{\hat{y}_t^{1st}}(h_t(g_t(x_t))) \\ &-\gamma\, \mathbb{E}_{(x_t,\hat{y}_t^{2nd})\in\mathcal{X}_t\times\hat{\mathcal{Y}}_t} \log \delta_{\hat{y}_t^{2nd}}(h_t(g_t(x_t))), \end{aligned} \tag{11}$$

where β and γ are trade-off hyperparameters which are set to 0.3 and 0.1 respectively.

Table 2. Top-3 Accuracies (%) on Asia test split and challenge test set of GeoPlaces challenge.

Method	Asia test split	Challenge (eval.ai)
Baseline	-	41.60
old source	66.36	50.24
old adapt	67.73	**51.23**
new source	66.52	50.56
new adapt	67.26	50.85

4.3. Result

We evaluate our solution on Asia test split and challenge test set of GeoPlaces challenge and results are shown in Table 2. Our solution achieves a top-3 accuracy of 51.23% on the challenge test set.

5. Challenge track: UDA on GeoImNet

5.1. Task

GeoImNet [1] is a challenging object classification benchmark whose images are collected from two continents with large geographical gaps. It contains two distant domains (i.e., USA and Asia) and 600 object categories. The task aims to utilize the annotation knowledge in the USA domain to improve the classification performance in the unlabeled images from Asia domain. The performance is evaluated by the top-3 accuracy.

5.2. Target Adaptation

In the target adaptation stage, the model is fine-tuned with unlabeled target dataset $\{x_t^i\}_{i=1}^{n_t}$. Following SHOT [2], we use the source model as initialization, freeze the classifier module h_t, and optimize the feature extractor g_t.

We employ the self-training strategy which learns representations under the supervision of the pseudo label. To obtain higher quality pseudo-labels, inspired by SHOT [2], we attain the centroid for k-th category in the feature space,

$$c_k = \frac{\sum_{x_t\in\mathcal{X}_t} \delta_k(h_s(g_s(x_t)))\, g_s(x_t)}{\sum_{x_t\in\mathcal{X}_t} \delta_k(h_s(g_s(x_t)))}. \tag{12}$$

Then, the pseudo labels are obtained based on the cosine distance between samples and centroids,

$$\hat{y}_t = \underset{k}{\text{argmax}}\ cosine(g_s(x_t), c_k). \tag{13}$$

Table 3. Top-3 Accuracies (%) on Asia test split and challenge test set of GeoImNet challenge.

Method	Asia test split	Challenge (eval.ai)
Baseline	-	49.65
old source	69.83	63.31
old adapt	70.61	**64.49**
new source	70.00	63.29
new adapt	70.85	64.46

Based on the pseudo label $\hat{y}_t$, the cross-entropy loss is calculated as,

$$\mathcal{L}_{tar}^{ImNet}(g_t) = -\eta \, \mathbb{E}_{(x_t,\hat{y}_t)\in\mathcal{X}_t\times\hat{\mathcal{Y}}_t} \log \delta_{\hat{y}_t}(h_t(g_t(x_t))), \tag{14}$$

where η is a hyper-parameter which is set to 0.3.

5.3. Result

We evaluate our solution on Asia test split and challenge test set of GeoImNet challenge and results are shown in Table 3. Our solution achieves a top-3 accuracy of 64.49% on the challenge test set.

6. Implementation Details

In this section, we provide the details of the whole process. We use Swin Transformer Base (Swin-B) [3] as the backbone in all experiments and we load ImageNet-1k pretrained parameters provided by Torchvision. A {Linear-BN-Linear} module is followed by Swin-B backbone and the total number of the parameters is 87,160,336 (less than 88M).

We adopt the learning rate scheduler $\eta = \eta_0 \cdot (1 + 10 \cdot p)^{-1.0}$, where p is the training progress changing from 0 to 1. The initial learning rate η_0 is 1e^{-3} for Swin-B backbone and 1e^{-2} for the rest in both the source training and target adaptation stage. We run source training for 10 epochs, while the maximum number of epochs during target adaptation for GeoUniDA, GeoPlaces, and GeoImNet is set to 5, 1, and 1, respectively. Besides, the batch size is set to 64 in all experiments.

References

[1] Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proc. CVPR*, pages 15368–15379, 2023. 2, 3

[2] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proc. ICML*, pages 6028–6039, 2020. 1, 2, 3

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. ICCV*, pages 10012–10022, 2021. 1, 4