ProDA

- Abstract
- 1. Introduction
  - 动机
    - 自训练是域自适应分割中的一种竞争方法。它使用目标域上的伪标签训练网络。然而不可避免的是，由于源域和目标域之间的差异，
  - 问题
    - 1. 伪标签是有噪声的。
    - 2. 并且目标特征是分散的。
  - 方法
    - 我们吸纳代表性原型，即类的特征质心，来解决无监督域适应的两个问题。
  - 具体
    - 特别地，我们更进一步，利用每原型的特征距离，提供比单点的原型更丰富的信息，具体来说。我们用它来估计伪标签的可能性，以促进训练过程中的在线修正。同时，我们根据同一目标的两个不同视图的相对特征距离对齐原型分配，产生更紧凑的目标特征空间。此外，我们发现将已经学过的知识提炼为自我监督的预训练模型进一步提高了性能。
  - 现状
    - 尽管深度学习在计算机视觉方面取得了巨大成功，但要获得足够数量的标签通常很昂贵。获取密集预测任务的标签通常很费时，例如语义分割。因此，人们考虑利用大量具有自由生成标签的道具合成图像（4,5,46）。然而，众所周知
  - 问题
    - 深度神经网络对域偏位非常敏感，渲染图像中任何细微的不真实都会导致对真实数据的泛化能力不佳。为了解决这个问题，域适应技术旨在以最小化性能损失将从合成图像（源域）学习的知识转移到真实图像（目标域
    - UDA
      - 无监督域适配（UDA），目标域中没有可访问的标签，真值未知，我们解决了语义分割的UDA问题
  - Self learning
    - self-training
    - 这是通过基于目标数据上最可靠的预测迭代生成一组伪标签。然后使用这些伪标签重新训练网络来实现的。这样，网络逐渐学会了在自定进度的课程学习中的适应性。然而，其性能仍然远落后于使用完全基于真实的标签基于监督学习。使用无监督策略来适应在实际场景中不实用
  - 问题
    - dissecting
    - 以前的工作中缺少两个关键因素。首先，典型实践(5,76)建议将最严格的置信阈值选择伪标签，而高分并不一定正确。这使得网络无法学习目标领域的可靠知识。其次，由于噪声鲁棒，网络容易在目标域中产生分散的特征，对于目标数据，相同记录分布，更容易评价模型。因此，在特定场景，将多这不会考虑这类原分布（即类分布）的数据
  - 方案
    - In this paper
    - 在本文中，我们提出在线去噪伪标签和学习一个紧凑的目标特征空间以解决上述两个问题。我们求助于原型，即类特征质心，来完成这两个任务。（1）我们根据伪标签与所有类原型的相对特征距离计算的可信度，从而对之伪标签，这取决于一个实际的假设，即原型更接近类真质心，这意味着错误伪标签少数，信用注意的是，原型是动态计算的，因此在整个训练过程中，伪标签会逐步得到纠正。（2）我们从Deepcluster(4)中得到启发，学习目标域的内在结构，而不是直接从属于分配中学习不同。我们建议为同一目标的不同视图调整软类别分配，从而产生更紧凑的目标特征空间。我们称本我们的方法ProDA，因为我们严重依赖于原域自适应的原型
    - 此外，我们发现将所已经学过的知识提炼为自我监督的预训练模型进一步提高了性能。
  - 贡献
    - contributions
    - 1. 我们的工作提供了一种新的无监督域自适应方法。
    - 2. 我们建议根据每原型的相对特征距离在线校正伪标签，并原型也会实时更新。因此，网络在整个训练过程中从去噪的伪标签中学习
    - 我们建议依靠每原型任务来教探索原型的学习，从而获得紧凑的目标特征空间
    - 将已经学习的知识提取到一个自我监督的预训练模型中，可以进一步显著提高性能
  - Unsupervised domain adaptation
    - 域对齐方法的重点是通过在图像级[1,12,13,20,27,50,64]、中间特征级[8,38,49,59]或输出级[35]上优化一些分歧(31,36)或采用对抗性训练21 41]来减少分布失配。然而，对于是分布不能保证目标域上的预测误差很小[7,29,72]。最近的方法[16,37,60]试图以类方式对齐分布，协同采用相关等特征对齐。**事实上，只要功能分离良好，就没有必要严格地对齐分布。**
  - 现有的解决方案
    - semi- supervised learning (SSL)
    - 另一方面，源于半监督学习（SSL）的技术提供了具有竞争力的性能。**佩森小化及其变体[9，47，56]提倡网络对未标记的目标数据进行熵最小化，由此产生的模型容易对噪声扰动不敏感。**
    - 解决这个问题
      - To address this, self- training
      - 迭代生成伪标签。然而，伪标签不可避免地有噪声。因此，[76]的网络中添加了重信正则项，而[74]则确信计了预测置信度，以减少不可靠标签的影响。在[33]中，自我训练和图像增强是互利的，最近的一项工作[60]基于来新图心生成伪标签，并在原始和增强数据上对齐。然而，这些由我训练方法是以另一种方式优化的——标签是通过学习过程中最显著的，只有在整个训练阶段能不会更新。
      - online pseudo label updating scheme
      - 标记三下，我们提出了一种在线修标签更新方案，根据目标域中在计的原型上下文标签重预测
  - 2. Related Work
  - Unsupervised representation learning.
    - 最近，无监督学习吸引了大量的研究兴趣[41]。早期的工作致力于设计**借口任务**[17,19,30,70]，当将其用作目标数据上的编码器时，有34.54 的的，证明 IID DA是有益的。一些基于相比学习的方法也著作[10,24]在大规模ImageNet上取得了与监督学习的差距。
    - consistent learning and cluster-based representation learn- ing
    - 最近的一系列工作[2,5,22,40]发现，网络聚类学习本重的语义特征。只要它们在不同的增强图下保持一致。然而，这些方法容易陷像区分[65]，因此不适合学习分割任务的像素级语义。**在这项工作中，我们发现一致性学习和基于聚类的表示学习的结合非常适合UDA问题，并从Deepcluster(4)中学习了一个紧凑的目标特征空间。不同的是，我们为不同的增强图片所每对特征距离，而不是聚类分配**
  - Learning from noisy labels.
    - 自我训练，即使小心减缓仍然会给噪声伪标签。因此，这项工作的动机也是从**噪声的标签中学习的新兴技术**[53]。
    - 解决方法
      - 一种简单的方法是设计**鲁棒损失**[63,71]。但这些方法无法在现实世界的噪声数据，**自我标签修正**[52,62,73,74]是一种更具吸引力的方法。该类别下的一种典型方式是[30]提用的协同网络**一起多个学习者**，并利用他们的预测一致性来衡量标签的可靠性。
      - Our proposed pseudo label denoising
      - 我们提出的伪标签去噪更接近于[30]。它根据一在复杂的**启发式算法确定的原型**在线标记**每原的标签**，概括之下，**我们能够实时计算原型**，另一方面，**知识提炼（KD）**[28,33,34]的建议明显有效的。可以将干净的知识从教师模型传递给学生，即使网络从自我学习(10,30)。在这项工作中，我们证明了自监督预训练模型的知识提取进一步推动了我们任务中的性能提升
- 3. Preliminary
  - aim to train a segmen- tation network
  - cannot generalize well to the target data due to the domain gap
- 4. Method
  - 4.1. Prototypical pseudo label denoising
    - 创新点来源
      - conjecture
      - 在一个训练阶段之后更新伪标签为时已晚，因为网络已经记忆和拟合了噪声标签。另一方面，当同时更新伪标签和网络权重时，容易陷入**简单的解决方案trivial**
    - 方案
      - a simple yet effective approach to online update the pseudo labels while avoiding trivial so- lutions
      - 关键
        - 关键是修复部伪标签，并按发现率逐步加权。并根据新学到的知识进行更新。形式上，我们建议使用指数移动平均来更新
        - 在线更新伪标签
          - $$q_i^{(t)} = \alpha q_i^{(t-1)} + (1-\alpha) p_i^{(t)} \quad (3)$$
          - 其中$q_i = q(x_i)$是你们要学习的伪标签，$p_i = p(x_i; \theta)$是当前网络的softmax输出。$\alpha \in [0, 1)$是指定动量的标量。
        - 权重更新公式
          - $$\alpha_t^{(c)} = \alpha \frac{\max_j \{\sum_{i=1}^{n} \mathbb{1}[y_i = c] \cdot p_j^{(t-1)}\}}{\sum_{i=1}^{n} \mathbb{1}[y_i = c] \cdot p_i^{(t-1)}} \quad (4)$$
          - 我们利用类原型的距离来估计正伪标签，$q(x) \in \{c\}$ 表示(x)是它的特征，如果它比其原型c_y，即c_y的特征更心。因此学习特征更有可能是一个非常准确的标签。除了原型可提供一个非常准确的概率，其余的是，我们将有它的得到**纠正**。**在这种情况下，我们将**α**设置为0.99**，**在整个学习过程中保持不变**，因此用作指数移动平均值
    - 原型计算初始化
      - Prototype computation.
      - $$c^{(t)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{1}[y_i^{(t)} = c] \cdot \mathbb{1}[q_i^{(t)} = j] \cdot z_i^{(t)}}{\sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{1}[q_i^{(t)} = j]} \quad (5)$$
      - $$c_j^{(t)} \leftarrow \lambda c_j^{(t-1)} + (1-\lambda) c_j^{(t)}, \quad (6)$$
      - 式中，$c_j^{(t)}$ (6) 是 j 动量编码器在当前训练批次内计算的类原型。λ 是预定义超参数，**我们设置为0.999的动量系数**
    - 伪标签训练损失
      - Pseudo label training loss
      - 通过在纠错伪标签，我们可以训练网络进行目标域分割。我们反传使用标准交叉熵，而是采用了更稳健的损失，即**广义交叉熵（GCE）损失**[55]。以进一步使模型**对噪声不敏感**，从而提高泛化能力。
      - $$\ell_{ce} = -\mathbb{1}[y_i = c] \cdot \log p_c(x_i; \theta) \quad (7)$$
      - 其中$p_c(x_i; \theta)$是你生模型的输出。我们设置$q = 0.7$，在实践中，一旦模型收敛，这种自信度**能以多次批**，这有助于稳定训练和实现更高的性能
    - Why are prototypes useful for label denoising
      - 首先，原型分配应为个别的实例提供平滑的目标。这可以通过将**一个实例的伪标签与其类原型进行插值来实现**。这可通过公式的形式来，因为未不同是一个具有挑战性的问题。
      - $$\ell_{ce} = \ell_{ce}(p_i, y_i) + \beta \ell_{ce}(p_i, c_{y_i}) \quad (8)$$
      - 其中$c_{y_i} \in \mathbb{R}^d$是你生模型的输出。我们设置$\beta = 1$。在实践中，一旦模型收敛，这种自信度**能以多次批**，这有助于稳定训练和实现更高的性能
    - 问题
      - a few pseudo labeled data cannot cover the entire dis- tribution in the target domain
      - 少数伪标记数据无法覆盖目标域中的整个分布
    - 目的
      - aim to learn the underlying structure of target domain, and hope to obtain more compact features
      - 为此，我们的目标是了解目标域的底层结构，并希望获得更紧凑的。有利于伪标签修改的特征。受最近无监督学习成功的激励，我们同时进行聚类和表征学习，与针对原型作业的学习不同，我们**使用我们扩充了的原型作业来指导深扩充视图的学习**。具体来说，让 $\{\mathbf{z}_i\}$ 和 $\{\mathbf{z}'_i\}$ 分别表示ent的目标"视图"和原标准"视图"。我们利用动量编码器为了 $\{\mathbf{z}'_i\}$ 生成一个可量的原型向量，它是
      - 弱视图特征权重
        - $$\alpha_i^{(t)} = \frac{\exp(-\frac{1}{\tau}\|\mathbf{z}'_i(t) - \mathbf{c}^{(t)}(y_i)\|_2^2)}{\sum_{j=1}^{K} \exp(-\frac{1}{\tau}\|\mathbf{z}'_i(t) - \mathbf{c}^{(t)}(j)\|_2^2)} \quad (9)$$
        - 其中τ默认为0.1。同样，设定$\alpha_i(t) \neq 0$ 对于 $\forall i$ {$\mathbf{z}'_i$}。**除了原始原型的不同权重外设置取值外，可以以相同的方式获得**
      - 弱视图指导强视图
        - $$\ell_{ce} = \mathbb{E}_i \ell_{ce}(\mathbf{z}_i, \mathbf{c}^{(t)}(y_i)) \quad (10)$$
        - 由于**此**可用，因为该特征由动量编码器给出，且输入vi的真标心，因此我们使用它来教的T {$\mathbf{z}'_i$} 生成一致的原型。因此，我们最小化了原种相同原型之间的Kullback-Leibler（KL）差异
        - 直观地说，**该方程强制网络为每张特征点提供一致的原型标记，从而使目标域中的特征空间更加紧凑**
    - 一个聚类为空
      - the proposed prototypical consistent learning may suffer from degeneration issue
      - 与以往同时学习聚类和表示的工作类似，本文提出的原型一致性学习可能存在退化问题，即一个聚类变为空。为了修正这一点，我们使用了[16]中的正则化
      - 正则
        - $$\ell_{reg} = -\frac{\alpha}{K} \sum_{j=1}^{K} \log p_j^{(t)} \quad (11)$$
      - 在线标签校正和一致性损失
        - $$\ell_{total} = \ell_{ce} + \ell_{ce}' + \gamma \ell_{ce}' + \gamma \ell_{reg} \quad (12)$$
        - 将在取各了在线标签校正和原型一致性学习，我们加除了各取以下总损失的加权组合用每
        - 默认情况下，加权系数γ和α分别设置为10和0.1
    - 收敛后蒸馏模型，
      - a student model
      - 在方程12的训练收敛后，我们进一步蒸馏**从学习的目标模型特殊预测是有相同结果结构的以及教室复杂式预测器的学生模型**。具体来说，我们使用**Res-CLIPv2**[11]预训练和更新的学生模型+的特征提取器，并在特别的动量编码器（MEE）的帮助下生成目。学生通过最小化更**其预测和更新的学生模型+的特征提取器**的交叉熵，**使用**[28]**的蒸馏损失**。此外，通过标准训练方式，我们将高价值的伪标签**用于学习一个紧凑的目标特征空间**。**不同的是，我们为不同的增强图片所每对特征距离，而不是聚类分配**
        - $$\ell_{kd} = \ell_{ce}^{s}(p_s, y_i) + \ell_{ce}^{s}(p_s(x_i), c_{y_i}) + \mathbf{KL}(p_t(x_i) \| p_s(x_i)) \quad (13)$$
        - 其中$p_t = \tau + \tau$（x）是学生模型的输出，我们设置$\tau = 1$。在实践中，一旦模型收敛，这种自信度**能以多次批**，这有助于稳定训练和实现更高的性能
  - 5. Experiments
    - 5. 1. Implementation
      - Training
      - 使用**DeepLab-v2**(8)作主干网**ResNet-101**(26)进行训练，初始化了(1)之后，我们利用SGD优化器训练网络，初始学习率为**1e-4**，每个训练用图像**裁剪为**，训练持续每个周期，在结束学习过程中，扩充出随机裁剪，随机扩**到512**和到**1140**两成，对于知识蒸馏，我们也使用了带有**ResNet-101**主干的预训练**DeepLab-v2**模型，在特征蒸馏之前引入一个额外的卷积模块（BN）层，以适应特征形状的差异指标，在**目标域**之前的之后，学习率分别设置为**6e-4和6e-3**
      - 在实验中，我们使用通用编造为**0.9**的动量指标，请参阅以参考以在的有了解更多有关指标和参量选择的，我们使用Pytorch在单台TESLA V100 GPU上进行所有实验
    - 5.2. Comparisons with state-of-the-art methods
      - Dataset
      - 为了评估，我们将数据集像素，GTA5(43)和SYNTHIA(44)数据集的分割模型为真实场景，城市景观(14)数据集，GTA5包含24640张分辨率为1914×1052的训练图像，我们使用与城市景观共享的19个类别。SYNTHIA数据集包含9400 1280×760幅图像，我们只用16个常见类别用于城市景观。我们也报告一些方法的结合，报告了消除**最高上限**之下常见类别的结果。Cityscapes数据集包含2975张训练图像和500张验证图像，分辨率为2048×1024。由于其测试集不提供地面真实标签，我们对其验证集进行评估
      - Table 1
        - 表1显示了GTA5的比较 → Cityscapes场景，我们的ProDA比其了最先进的mIoU分数**7.5**，大大超过了现有方法。在所有19个类别中，**我们在地个类别中得分最高**。ProDA在绝等每种类三分别取胜的任务，例如**摩托**、**平台**、**电车**。这些在以前的工作中无法获得胜利**。事实上，ProDA的性能改进主要来自这些具有挑战性的案例，即小型或稀有物品。因为我们对不同类别的产品同样重视，这要归功于原型**。注意，这是在没有任何经验或不平衡数的情况下实现的。**即**70%**，**节省食品后果将（即，种样在源上训练的模型）**。**特别**，ProDA是你的mIoU增益为20.9，低于次优方法8.6**
      - In Table 2,
        - 在表2中，我们展示了SYNTHIA场景的结果 → 城市景观。**整**的结论是示出巨大的改进，建议的ProDA比*的7和13个**类中的mIoU分数分别达到65.5和40.6。**此外**此**，我们在16个类别中的**11个类**上**取得了最好的成绩**。其中大部分是高难度课程，**与带有语义模型相比，自适应的分割模型获得了8.6%的改**进，比先前的最先方法高出了**58.4%**。虽然SYNTHIA场景比GTA5更具挑战性，但我们的方法在两个数据集上都表现良好
      - The effectiveness of pseudo label denoising
        - 在表11中，带有蒸馏原模型在目标域上获得36.6 miou，在**蒸馏预测后**，我们得到了5.0miou的增长。通过预热初始化，基线模型，即使用最有信与标签训练的意果/普通最新训练，在miou提高到42.2，高出**预设交换（SCE）**可增长+0.4 miou 的增益，相比之下，我们即使在更新各标签，使用**动量**标签进行训练可显著提高性能至52.3。仅使用此程序的**蒸馏**就达到了最先进的水平，比之最好最佳分数高出2.0
      - Figure 2
        - 说明了伪标签在培训过程中的进度。在早期阶段，可能会出现虚假的伪标签，无法识别图像小物体，例如电线杆和交通标志。随着训练的进行，伪标签会细化原型方案，并且可以在最长的伪标签中选择识别出微小的对象。经过**40k次迭代**后的情况**证明**可以标记大多数**不正确的标签**。在图3中，我们进一步量化了**整个**培训过程中伪标签的mIoU计算。虽然常规会随培训保持每个**的**标记更新伪标签，并导致准确**miou**增加，**但**ProDA**提供的**标签可以快速获得高质量**。与多个训练阶段的普通自我训练相比，表现出明显的优势
    - 5.3. Discussion
      - How to prevent degenerate solution?
        - learning the feature and updating the labels during self-training derives degenerate solutions
      - The effectiveness of target structure learning
        - 目标底层结构
          - 我们建议为目标domain推荐内在知识，并**学习紧凑表格**。如果1)和2)，**在没有特定结构的情况下，使用学习原性能达45.4 miou高到47.6**。注意，这也是所有训练方法之间的竞争。但我们不需要行相选择和最本质的原性特征，**目标结构学习通过取成紧凑的特征来辅助伪标签去噪，使性能提高1.4%**
      - The effectiveness of distilling to self-supervised model
        - At the second training stage
        - 在第二个训练阶段，我们应用知识提取，并用第一阶段模型的相知识将标签进行**蒸馏**。**在**第二**阶段**，我们**使用**在第一阶段**训练模型**，**作为**教师模型（即**在**使用**ResNet-101**主干的预训练模型）**，**学生模型**（即**在**使用**ResNet-101**主干的预训练模型）**，**学生模型**通过**蒸馏**学习**。**在**第二阶段**，**学生模型**的**性能**进一步提高了**6.6**。**与**在**每**个**阶段**的**性能**进一步提高了**10.9 miou**。**在**蒸馏**阶段**，**学生模型**的**性能**甚至**超过了**教师模型**，**性能提高了20.9 miou**
      - The UMAP visualization of target features.
        - visualize the learned features
        - 为了更好地直观直观，我们在图3中可视化了ProDA的学习过程。随着起初的模型原始了用一类的特征，伪标的非训练可以产生**更多的分离特征空间**。但对于**相性分类仍然是困难的**，当我们使用**蒸馏**伪标签去**精**时，不同类别之间的特征得到了**更好的分离**。因**此**伪标的质量分数的，相比之下，**我们的模型能**给了**最紧凑的特征集**，可以进行分类
        -
    - 5.4. Parameter sensitivity analysis.
      - robust
        - Table 3. The effect of threshold in pseudo label selection during the pseudo-label label denoising.

threshold	0.001	0.01	0.05	0.09	0.5	0.9
mIoU	51.7	52.0	52.3	52.1	51.6	51.2

          Table 4. The effect of prototype momentum during the pseudo-label label denoising.

momentum	0.001	0.900	0.990	0.9999
mIoU	51.4	51.6	52.3	50.1

          Table 5. The effect of prototype momentum during the pseudo-label label denoising.

          为了展示ProDA对超参数选择的**鲁棒性**，我们分析了参数的影响。在表4中，我们使用不同的阈值来选择伪标签，并且性能对**选择阈值**不**敏感**。因此，为了方便使用，ProDA不应用阈值，我们进即在了在每个类别**计算原型的质量**的**距离**，**并且ProDA对**"**这的微值范围**具有鲁棒性**，如表5所示
  - Conclusions
    - proto- types to online denoise the pseudo labels and learn the com- pact feature space for the target domain. Knowledge distil- lation to a self-supervised pretrained model further boosts the performance
    - 在本文中，我们提出了ProDA，以增强对原型的标签噪声的学习并学习目标域的紧凑特征空间。我们提倡以原型的动态信息对伪标签进行在线修正，并通过将知识提取到一个自我监督的预训练模型中进一步提高了性能。我们证明大大优于现有的方法，大大缩小了与监督学习的差距。我们的方法可以轻松地推广到其他任务，如实例分割(25)和一般领域适应。我们希望能将简单的原则应用于解决它的能力