**Modeling and Simulation** 建模与仿真, 2023, 12(3), 3051-3062
Published Online May 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.123281

Hans汉斯

# 基于Swin Transformer的无监督域自适应图像分类

**范博文，徐志洁***

北京建筑大学理学院，北京

收稿日期：2023年4月12日；录用日期：2023年5月24日；发布日期：2023年5月31日

## 摘　要

**大多数当前的无监督域自适应(UDA)技术从域级别或类级别学习域不变的特征表示。基于域级别的主流方法是对抗学习，对抗学习通常不考虑目标数据的固有判别信息。基于类别级别的UDA方法通常是为目标域样本生成伪标签，由于这些伪标签通常噪声太大，这不可避免地会影响UDA性能；其次，现有方法没有明确地强制区分不同类别的特征。为了解决以上问题，我们提出了基于Swin Transformer的无监督域自适应(SwinUDA)。首先，对于域对齐，将Swin Transformer与对抗性自适应相结合，提高模型对噪声输入的鲁棒性，其次，对于类别对齐，使用正交投影损失(OPL)直接在特征空间中实施约束。此外，正交投影损失对标签噪声干扰的影响更有鲁棒性。最后，引入了互信息最大化损失(IML)来保留目标域的可区分特征。本文提出的SwinUDA模型可以同时学习可迁移和可区分的特征。在Office-Home、Office-31和VisDA-2017三个公开数据集上进行实验，SwinUDA都展现了最佳的性能。**

## 关键词

**对抗学习，无监督域适应，图像分类，类别对齐，伪标签生成**

# Unsupervised Domain Adaptation Image Classification Based on Swin Transformer

**Bowen Fan, Zhijie Xu***

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Apr. 12th, 2023; accepted: May 24th, 2023; published: May 31st, 2023

## Abstract

**Most current unsupervised domain adaptation (UDA) techniques learn domain invariant feature**

*通讯作者。

文章引用: 范博文, 徐志洁. 基于 Swin Transformer 的无监督域自适应图像分类[J]. 建模与仿真, 2023, 12(3): 3051-3062.
DOI: 10.12677/mos.2023.123281

**representations from the domain-level or class-level. Adversarial learning is the dominating strategy based on the domain-level. It tries to align the global feature distributions of the two domains without considering the target data's innate discriminative information. Class-level-based approaches typically generate pseudo-labels for data in the target domain. These pseudo-labels impact UDA's performance because they are generally overly noisy. In addition, existing methods do not explicitly enforce a good separation of different classes of features. To solve the above problems, we propose the Unsupervised Domain Adaptation Using Swin Transformer (SwinUDA). First, for domain alignment, the Swin Transformer is combined with adversarial adaptation to improve the robustness of the model to noisy inputs. The experimental results show that using the transformer as a feature extractor has higher transferability. Second, constraints are directly enforced in the feature space for class alignment using Orthogonal Projection Loss (OPL). Samples from the same class (whether from the source or target domain) are pulled closer, while samples from different classes are pushed away. In addition, the orthogonal projection loss is more robust to the influence of label noise interference. To preserve the discriminative information of the target domain, a mutual information maximization loss (IML) is introduced to protect the discriminating features of the target domain. The SwinUDA model proposed in this paper can simultaneously learn transferable and differentiable features. Experiments were performed on the three public datasets Office-Home, Office-31, and VisDA-2017. SwinUDA showed the best performance.**

## Keywords

**Adversarial Learning, Unsupervised Domain Adaptation, Image Classification, Class Alignment, Pseudo-Label Generation**



## 1. 引言

深度神经网络在许多计算机视觉任务中取得了令人印象深刻的性能。然而，这些方法成功通常依赖于大量标记的数据，而获取这些数据非常的耗时并且获取成本高昂，此外，由于计算机视觉中光照、背景、天气条件等因素的影响，也经常会出现训练数据和测试数据之间的特征分布不匹配。因此推动了无监督域适应(UDA)的研究，UDA 任务旨在存在域偏移的情况下，将知识从标记的源域转移到不同的未标记目标域。

大多数当前的方法[1] [2] [3] [4]试图通过对齐两个域的特征分布来学习域不变表示。利用生成对抗网络(GAN)的思想是一种常见的技术[5]。通过生成器和判别器之间的极小极大博弈进行模型训练。然而，如果强制将两个域中特征的全局边缘分布对齐，可能会忽略每个类别的局部联合分布。这种忽略会导致目标域中原本已经对齐的类别经过训练后映射到错误的类别，如图 1 所示。另一种流行的方法旨在进行类别级的对齐来学习目标域的区分特征[6]-[11]。基于类级对齐的主要方法是生成与目标样本概率匹配的伪标签，并使用这些伪标签来训练模型。然而，这些伪标签通常噪声太大，无法进行精确的域对齐导致模型性能下降[10]。同时没有强制不同类别的特征很好的分离。

总之，域级对齐可以对齐源域和目标域的全局特征分布，以学习可迁移特征。而类别级对齐可以学习有区别的目标特征。理想的方法是结合这两种方法的优点，同时强制不同类别的特征分离。为了实现这一目标，我们提出了一种新的 UDA 解决方案，即 SwinUDA (基于 Swin Transformer 的无监督域自适应)。

首先，通过将 Swin Transformer 与简单的对抗性域适应相结合，以对齐源域和目标域的全局特征分布。实验结果表明，Swin Transformer 具有很强的可迁移性。为了更好地区分不同类别的样本，我们同时考虑类别级的对齐。引入了正交投影损失(OPL)，并使用伪标签辅助计算该损失。OPL 可以强制将相同的类特征很好地聚类，将不同的类特征很好地分离。为了保持目标域的内在结构，引入了互信息最大化损失(IML)，以保留更多的目标域信息，并进一步提高模型性能。

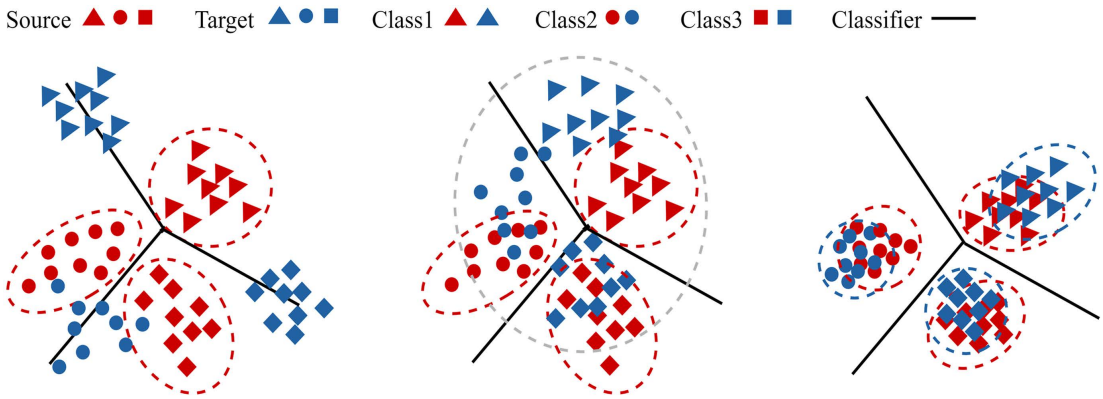



**Figure 1.** (Best viewed in color) Contrast our suggested approach with earlier research. Left: Trained using only data from the source domain, applied directly to the target domain. Middle: Globally aligns the data distributions of the two domains, regardless of class information. Right: Our proposed method considers class-level domain alignment, reducing conditional distribution differences

**图 1.** (最好用彩色观看)将我们的方法与早期的研究进行对比。左图：仅使用源域中的数据进行训练，直接应用于目标域。中间：全局对齐两个域的数据分布，而不考虑类信息。右图：我们提出的方法考虑了类级别的域对齐，减少了条件分布的差异

本文的主要贡献是：1) 我们提出的 SwinUDA，是第一次将 Swin Transformer 作为无监督跨域图像分类的主干网络，为了保护目标域的内在结构，我们引入互信息最大化损失，以减轻对抗域适应中的目标域区分破坏。2) 为了使同类特征接近，不同类特征分离，我们引入 OPL 损失，同时学习可迁移特征和可区分特征。3) 在 Office Home、Office-31 和 VisDA-2017 三个公开数据集上的实验表明，我们的 OPST 都展现了最佳的性能，其中，Office Home 为 87.17%，office-31 为 94.6%，VisDA-2017 为 88.46%。

## 2. 相关知识

### 2.1. 问题设置

无监督域适应的目标是处理来自 $\mathcal{X}\times\mathcal{Y}$ 的有标记的源域数据 $D_S=\left\{\left(x_i^s,y_i^s\right)\right\}_{i=1}^{n_s}$ 和来自 $\mathcal{X}$ 的未标记的目标域数据 $D_T=\left\{\left(x_i^t\right)\right\}_{i=1}^{n_t}$ 之间的域偏移问题，其中 $\mathcal{X}$ 是输入空间，$\mathcal{Y}$ 是标签空间，$n_s$ 和 $n_t$ 分别为源域和目标域的样本数量。假设它们假设特征空间、标签空间与条件概率分布都相同，即 $\mathcal{X}_s=\mathcal{Y}_t=\mathbb{R}^d$，$\mathcal{Y}_s=\mathcal{Y}_t=\left\{y^1,y^2,\cdots,y^C\right\}$，$P\left(y^s\middle|x^s\right)=P\left(y^t\middle|x^t\right)$。但这两个域的边缘分布不同，即 $P\left(x^s\right)\neq P\left(x^t\right)$。UDA 的任务是利用有标签的源域数据学习一个分类器 $h=g\circ f$ 来预测目标域数据 $D_T$ 的标签 $y^t\in\mathcal{Y}_t$，其中 $f\left(\cdot;\theta_f\right):\mathcal{X}\to\mathcal{Z}$ 表示特征提取器，$g\left(\cdot;\theta_g\right):\mathcal{Z}\to\mathcal{Y}$ 表示类别预测器，$\mathcal{Z}$ 表示特征空间。

### 2.2. Swin Transformer 模型

目前，transformer 在计算机视觉领域的应用面临两个局限：第一，视觉目标大，视觉 transformer 在不同场景下的性能较差；第二，当图像分辨率高时，transformer 的计算量大。为了解决上述两个问题，

Swin transformer [11]提出了一种滑动窗口操作，该操作以分层方式构建 transformer，并将注意力计算限制在一个窗口内，这大大减少了计算量。

Swin transformer 模型如图 2 所示。首先，根据 4×4 个相邻像素将输入图像划分为一个 patch，并通过 patch 划分将每个 patch 在通道方向上展平。其次，堆叠 4 个 stage 来构建不同大小的特征图，用于注意力计算。每个 stage 代表一个层次。第一个 stage 通过线性嵌入改变特征维度，最后三个 stage 通过 patch merging 进行下采样并重复堆叠 Swin transformer block。多层感知机、窗口多头自注意层、滑动窗口多头自注意力层和标准化层构成了 Swin transformer 块的大部分，如图 2 右侧所示。其中，自注意力层是 transformer 的关键组件，其计算方法如下式所示：

$$Attention(Q,K,V)=soft\max\left(QK^T/\sqrt{d}\right)V \tag{1}$$

其中 $Q$，$K$，$V$ 分别为 query、key、value，$d$ 为查询维度。Transformer 中的注意力机制对噪声输入具有鲁棒性，可以更好地提取信息全局特征。

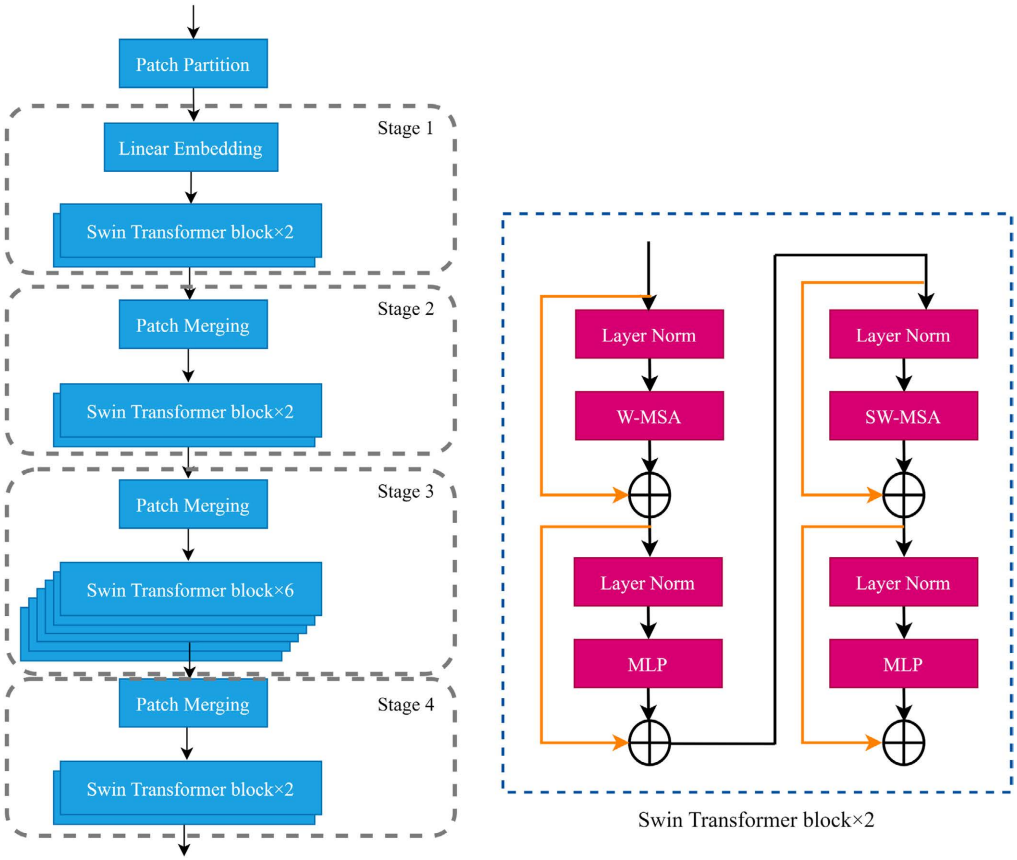



**Figure 2.** Swin Transformer architecture
**图 2.** Swin Transformer 网络架构

## 3. 方法

### 3.1. 网络架构

提出的网络结构如图 3 所示。它由三部分组成：特征提取器(SwinT)、标签分类器(Label Classifier)和领域判别器(Domain Discriminator)。对于每个源域和目标域图像，通过一系列 transformer blocks 提取特

征。域判别器判断输入图像是源样本还是目标样本。域判别器的训练目标是将输入尽量分到正确的域，而特征提取器所提取的特征目的是使域判别器不能正确的判断出信息来自哪一个域。以这种对抗性的方式训练域鉴别器和特征提取器以进行域对齐。标签分类器获得类标记并输出标签预测，通过计算目标样本的互信息最大化损失(IML)来减轻对抗性学习中目标样本的结构损坏。同时，正交投影损失(OPL)可以使样本在特征空间中实施正交约束，实现类内特征聚类和类间特征分离。

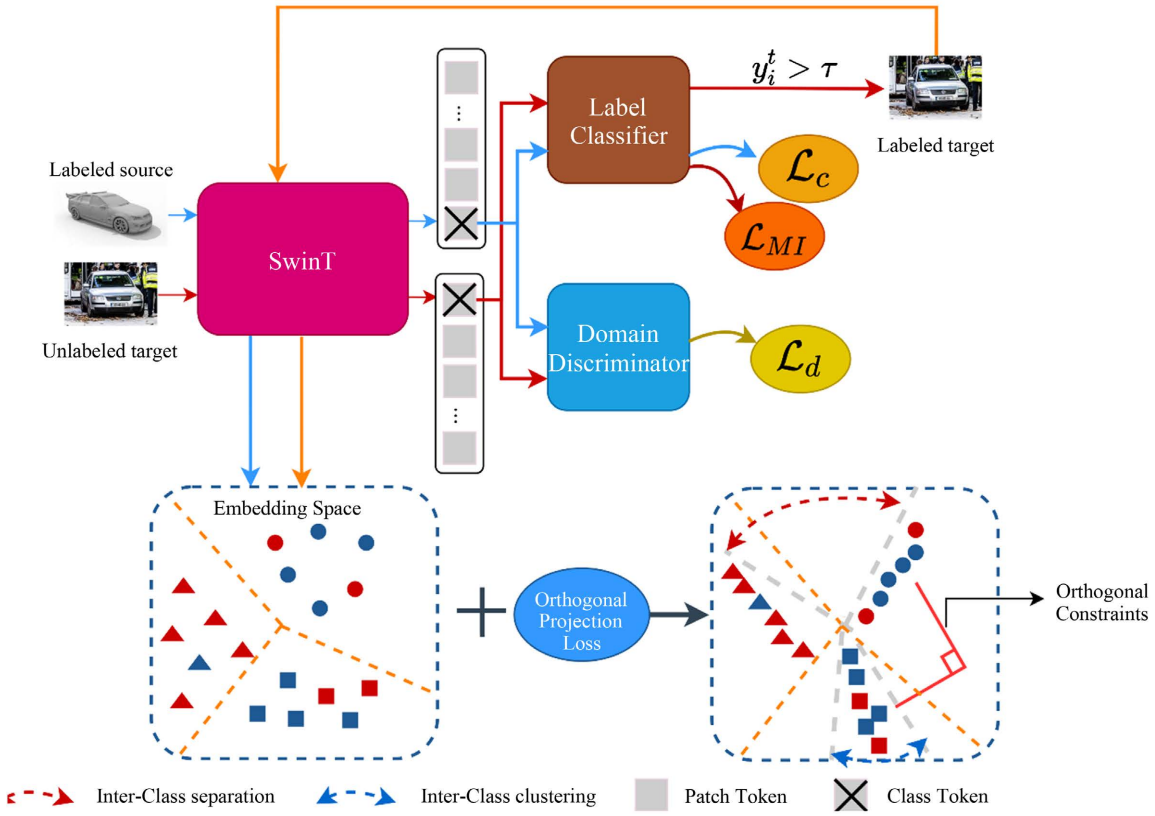



**Figure 3.** Network architecture
**图 3.** 网络架构

## 3.2. Swin Transformer 对抗域适应

遵循典型的对抗性自适应方法来实现领域自适应。旨在利用有标签的源域数据 $D_S = \left\{\left(x_i^s, y_i^s\right)\right\}_{i=1}^{n_s}$ 学习一个分类器 $h = g \circ f$ 来预测目标域数据 $D_T = \left\{\left(x_i^t\right)\right\}_{i=1}^{n_t}$ 的标签 $y^t \in \mathcal{Y}_t$。其中 $f\left(\cdot;\theta_f\right): \mathcal{X} \to \mathcal{Z}$ 表示特征提取器，本章使用 Swin Transformer 作为特征提取器 $f$，$g\left(\cdot;\theta_g\right): \mathcal{Z} \to \mathcal{Y}$ 表示类别预测器，$\mathcal{Z}$ 表示特征空间。采用对抗学习的思想，添加一个域判别器 $d\left(\cdot;\theta_d\right): \mathcal{Z} \to [0,1]$，其中 $d$ 指示输入样本来自源域或者目标域。形式上，我们的对抗学习的目标是优化公式(2)来共同优化目标分类损失 $\mathcal{L}_c$ 和域对抗损失 $\mathcal{L}_d$：

$$\min_{\theta_g,\theta_f} \max_{\theta_d} \mathcal{L} = \sum_{i=1}^{n_s} \mathcal{L}_c\left(G\left(g\left(x_i^s;\theta_g\right);\theta_f\right), y_i^s\right) - \mathcal{L}_d\left(\theta_g, \theta_d\right) \tag{2}$$

其中，$\mathcal{L}_c$ 是源域数据的标准交叉熵损失，$\mathcal{L}_d$ 是域对抗损失，定义为：

$$\mathcal{L}_d\left(\theta_g, \theta_d\right) = E_{x\sim D_s}\left[\log d\left(g\left(x\right)\right)\right] + E_{x\sim D_t}\left[\log d\left(1 - g\left(x\right)\right)\right] \tag{3}$$

### 3.3. 损失函数

对于 UDA 中的域对齐问题，在对齐两域的全局分布的同时，还要尽可能的减轻目标域内在结构的破坏，因此要考虑一个问题，理想的目标输出是什么样子？我们认为完美的目标输出应该满足以下几点：1) 决策边界位于低密度区域，也称为聚类假设[12]。2) 防止所有目标数据被分类到同一类中。在许多领域自适应工作中，信息熵被用来最小化学习目标数据的区别特征。信息熵可以估计模型预测的不确定性程度。预测结果的准确性随着信息熵的减小而增加。标准信息熵计算方法为：

$$H\left(x^t\right)=-\sum p\left(x_i^t\right)\log p\left(x_i^t\right) \tag{4}$$

经过分析发现，标准的交叉熵计算方法无法准确评估样本伪标签在决策边界处的不确定性。而互信息最大化损失可以避免将所有目标样本分配给同一类，这满足了理想的目标输出。互信息最大化损失被证明比先前领域自适应工作中常用的信息熵最小化更有效[13]。

为此，采用互信息最大化损失：

$$\begin{aligned}\mathcal{L}_{MI}&=\mathcal{I}\left(p\left(x^t\right);x^t\right)=H\left(\left[\bar{p}\left(x^t\right)\right]\right)-\frac{1}{n_t}\sum_{i=1}^{n_t}H\left(p\left(x_i^t\right)\right)\\&=-\sum_{c=1}^{C}\bar{p}\left(x_c^t\right)\log\left(\bar{p}\left(x_c^t\right)\right)+\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{c=1}^{C}p\left(x_{ic}^t\right)\log\left(p\left(x_{ic}^t\right)\right)\end{aligned} \tag{5}$$

其中，$p\left(x_i^t\right)=soft\max\left(g\left(f\left(x_i^t\right)\right)\right)$，$\bar{p}\left(x^t\right)=\mathbb{E}_{x_t}\left[p\left(x^t\right)\right]$表示整个目标样本的平均输出嵌入，$\mathbb{E}_{x_t}\left[p\left(x^t\right)\right]=\frac{1}{n_t}\sum_{i=1}^{n_t}p\left(x^t\right)$表示目标样本的期望。最小化第二项可以导致目标预测接近一个热编码，而最大化第一项可以防止所有目标数据被放在同一类中。使用互信息最大化损失鼓励模型学习均匀分布的紧密的目标特征以便保留关于目标数据的更多判别信息。

目标是在全局对齐期间执行更好的域自适应并确保准确的类级对齐，同时在不同类的特征远离时保持同一类的特征接近。因此，在特征空间中实现了正交约束。给定来自数据集 $D$ 的标记样本$\left\{x_i,y_i\right\}$，$F_i=g\left(x_i\right)$是网络提取的特征，通过聚类$F_i$，使得不同类别的特征应尽可能正交，同一类的特征应尽可能相接近。因此，我们通过引入正交投影损失(OPL) [14]，以确保类内聚类和类间正交性：

$$\mathcal{L}_{OP}=\left(1-s\right)+\left|d\right| \tag{6}$$

其中：

$$s=\sum_{i=1}^{n}\sum_{j=1}^{n}CS\left(f_i,f_j\right),\ y_i=y_j \tag{7}$$

$$d=\sum_{i=1}^{n}\sum_{k=1}^{n}CS\left(f_i,f_k\right),\ y_i\neq y_k \tag{8}$$

其中$CS\left(\cdot,\cdot\right)$表示两个向量的余弦相似函数，$\left|\cdot\right|$是取其绝对值，要注意的是公式(7)和公式(8)中的余弦相似函数涉及特征归一化：

$$CS\left(x_i,x_j\right)=\frac{x_i\cdot x_j}{\left\|x_i\right\|_2\cdot\left\|x_j\right\|_2} \tag{9}$$

公式(7)和公式(8)定义了两个类差异，$s$ 测量类内特征差异，$d$ 测量类间特征差异。通过使 $s$ 接近 1，$d$ 接近 0 来最小化公式(6)，以实现最小化类内特征差异和最大化类间特征间距，无论样本来自哪个域。由于 OPL 的计算需要获得目标样本的标签，因此这里我们使用一种简单有效的方法来获得带伪标签的目

标样本。我们根据分类器的预测概率来选择目标样本，让 $\left\{p_c\left(x_i^t\right)\middle|_{c=1}^{C}\right\}$ 表示分类器的 softmax 层的输出，其中 $p_c\left(x_i^t\right)$ 表示样本 $x_i^t$ 属于第 $c$ 类的概率，$C$ 是类别总数。

然后可以得到目标样本的伪标签 $y_i^t=\arg\max_c p_c\left(x_i^t\right)$，称 $p_{y_i^t}\left(x_i^t\right)$ 为分类置信度得分。通过选择分类置信度得分高于阈值 $\tau$ 的目标样本，获得一个带有伪标签的目标样本集 $\tilde{D}_T=\left\{\left(x_i^t,\tilde{y}_i^t\right)\right\}_{i=1}^{n_t'}$，可以由标记的源域样本集 $D_S=\left\{\left(x_i^s,y_i^s\right)\right\}_{i=1}^{n_s}$ 和带伪标签的目标域样本集 $D_T=\left\{\left(x_i^t,\tilde{y}_i^t\right)\right\}_{i=1}^{n_s}$ 得到有标签的样本集 $D=\left\{\left(x_i,y_i\right)\right\}_{i=1}^{n}$，其中 $n=n_s+n_t'$。

简单地最大化类边界可能会导致类之间的负相关性，从而导致过分地关注分离良好的类，不好区分的困难样本被忽视。而正交投影损失倾向于确保不同类特征之间的独立性，以成功地分离类特定特征。尽管目标域的伪标签估计可能有噪声，但正交投影损失在一定程度上对噪声有鲁棒性，能够减少伪标记的错误影响。

因此训练总目标为：

$$\mathcal{L}_c\left(x^s,y^s\right)+\mathcal{L}_d\left(x^s,x^t\right)+\alpha\mathcal{L}_{IM}\left(p^t,x^t\right)+\beta\mathcal{L}_{OP}\left(x^s,x^t\right) \tag{10}$$

其中 $\alpha$ 和 $\beta$ 是超参数。

## 4. 实验结果与分析

### 4.1. 实验设置

一个 bottleneck 模块(Linear → BatchNorm1d → ReLU → Dropout(0.5))和标签预测器(Linear → ReLU → Dropout(0.5) → Linear)组成分类器头部。除了具有单个输出之外，域判别器还与标签预测器共享相同的网络结构。在训练过程中，首先将图像大小调整为 256 × 256，然后随机水平翻转，然后随机裁剪并调整图像大小至 254 × 254，唯一的变化是，在 VisDA-2017 [15]数据集中，使用了大小为 254 × 254 的中心裁剪。在测试过程中，首先将图像的大小调整为 256 × 256，然后中心裁剪到 254 × 254 像素。为了训练模型，使用动量为 0.9 的小批量随机梯度下降(SGD)。学习率设置为 $lr=lr_0*\left(1+1e^{-3}\cdot i\right)^{-0.75}$，其中 $lr_0$ 表示初始学习率，并且 $i$ 是训练步骤。特征提取器骨干参数的学习率被设置为 $lr$ 的 1/10。在所有的实验中设置 $\tau=0.8,\alpha=0.1,\beta=1.0$。

比较的方法包括基于 CNN 的方法 ALDA [16]、TADA [17]、SHOT [18]、MCD [19]、CaCo [20]、STAR [21]和基于 transformer 的方法 CDTrans [11]、TVT [22]、SSRT [23]。对于所有上述方法，在其原始论文中总结了报告的结果。“Baseline”是具有对抗性适应的 SwinT-B，“SwinUDA”是本文提出的方法。

### 4.2. 实验结果与分析

将对抗性适应的 Swin Transformer 作为基准模型(Baseline)，与基于 CNN 的域适应技术相比，优点表现在三个方面：首先，注意力权重和图像内容依赖于内容进行交互，这些交互可以被认为是空间变化的卷积。其次，通过滑动窗口机制实现了长程依赖建模。最后，它结合了 CNN 和 transformer 的优势，展现了卓越的潜力。如表 1、表 2 和表 3 所示，基准模型的绝对精度可以实现与现有技术的基于 CNN 的方法相当的性能。与目前性能最好的 SHOT 相比，它在 Office-Home 数据集[24]上提升了 11.2%，在 Office-31 数据集[25]上提高了 5.0%。这一结果表明了 Swin Transformer 的强大的可迁移性，并表明注意力机制和视觉内容之间的交互可以更好地收集特征信息。然而，在具有显著领域差异的 VisDA-2017 数据集[15]上，基准模型还有待改进，其一是该模型在确保全局分布一致的同时，不能保证边缘分布的一致性。其二是对抗性学习会忽略目标样本的固有信息，将不同类别的样本混合在一起，导致分类错误。

**Table 1.** Accuracies (%) on Office-Home. CDTrans* uses DeiT-base backbone. TVT* uses ViT-base backbone
**表 1.** Office-Home 数据集的精度(%)。CDTrans*使用 DeiT 基础骨干网。TVT*使用 ViT 基本骨干网

| Domains | ALDA [15] | TADA [16] | SHOT [17] | CDTrans* [11] | TVT* [21] | SSRT [22] | Baseline | SwinUDA |
|---|---|---|---|---|---|---|---|---|
| Ar → Cl | 53.7 | 53.1 | 57.1 | 68.8 | 74.89 | 75.17 | 73.91 | **77.80** |
| Ar → Pr | 70.1 | 72.3 | 78.1 | 85.0 | 86.82 | 88.98 | 86.66 | **91.19** |
| Ar → Rw | 76.4 | 77.2 | 81.5 | 86.9 | 89.47 | 91.09 | 88.80 | **91.85** |
| Cl → Ar | 60.2 | 59.1 | 68.0 | 81.5 | 82.78 | 85.13 | 81.54 | **86.81** |
| Cl → Pr | 72.6 | 71.2 | 78.2 | 87.1 | 87.95 | 88.29 | 84.59 | **90.52** |
| Cl → Rw | 71.5 | 72.1 | 78.1 | 87.3 | 88.27 | 89.95 | 85.88 | **91.05** |
| Pr → Ar | 56.8 | 59.7 | 67.4 | 79.6 | 79.81 | 85.04 | 81.62 | **86.61** |
| Pr → Cl | 51.9 | 53.1 | 54.9 | 63.3 | 71.94 | 74.23 | 73.10 | **78.21** |
| Pr → Rw | 77.1 | 78.4 | 82.2 | 88.2 | 90.13 | 91.26 | 89.76 | **93.25** |
| Rw → Ar | 70.2 | 72.4 | 73.3 | 82.0 | 85.46 | 85.70 | 85.37 | **86.81** |
| Rw → Cl | 56.3 | 60.0 | 58.8 | 66.0 | 74.62 | **78.58** | 75.53 | 78.03 |
| Rw → Pr | 82.1 | 82.9 | 84.3 | 90.6 | 90.56 | 91.78 | 91.20 | **93.85** |
| Avg | 66.6 | 67.6 | 71.8 | 80.5 | 83.56 | 85.43 | 83.16 | **87.17** |

我们在中等规模的 Office-Home 数据集上进行了对比实验，其结果如表 1 所示，提出的方法大大优于基于 CNN 的顶级无监督域适应技术 SHOT (87.17% vs. 71.8%)。可以观察到：当基于 CNN 的无监督域适应方法 TADA 仅考虑域对齐时，模型的性能相对较差。相比之下，SHOT 的类级对齐模型显示了显著的改进，证明了类级对齐对领域自适应至关重要。并且与基于 CNN 的无监督域适应方法相比，基于 transformer 的无监督域适应方法有了进一步的改进。与 SHOT 相比，考虑类级对齐的 CDtrans 方法有了显著的改进(从 80.5%提高到 71.8%)，这表明 transformer 在特征提取方面是强大的。此外，提出的方法使用 Swin Transformer 进行特征提取，考虑域级别和类级别的对齐，并产生最佳结果。同时，在具有显著的域偏移的 Ar → Rw 和 Cr → Rw 任务方面的表现优于 SHOT，表明 SwinUDA 在从具有挑战性的域移动到简单域时具有出色的鲁棒性和泛化能力。

**Table 2.** Accuracies (%) on Office-31
**表 2.** Office-31 数据集的精度(%)

| Method | A → W | D → W | W → D | A → D | D → A | W → A | Avg |
|---|---|---|---|---|---|---|---|
| TADA [17] | 94.3 | 98.7 | 99.8 | 91.6 | 72.9 | 73.0 | 88.4 |
| SHOT [18] | 90.1 | 98.4 | 99.9 | 94.0 | 74.7 | 74.3 | 88.6 |
| ALDA [16] | 95.6 | 97.7 | 100.0 | 94.0 | 72.2 | 72.5 | 88.7 |
| CDTrans* [11] | 96.7 | 99 | 100.0 | 97.0 | 81.1 | 81.9 | 92.6 |
| TVT* [22] | 96.4 | **99.4** | 100.0 | 96.4 | 84.9 | **86.1** | 93.8 |
| SSRT [23] | 97.7 | 99.2 | 100.0 | 98.6 | 83.5 | 82.2 | 93.5 |
| Baseline | 98.4 | 99.3 | 100.0 | 98.2 | 85.5 | 85.3 | 94.4 |
| SwinUDA | **99.1** | 99.3 | **100.0** | **984** | **85.8** | 85.0 | **94.6** |

为了进一步验证模型的有效性，我们在 Office-31 数据集上进行了对比实验，其结果如表 2 所示。提出的 SwinUDA 总体上优于所有对比的方法，并将最先进的结果平均从 93.5%提高到 94.6%。尤其是在具有挑战性的转移任务(如 A → W 和 A → D)，SwinUDA 也显示出显著的改善。与类级对齐方法 ALDA 和 SHOT、域对齐方法 TADA、、基于注意力机制的方法 TADA 相比，所提出的方法优于它们的性能可以表明 SwinUDA 的每个组件的有效性。上述结果证明，IML 的使用可以减轻对目标域内在结构的破坏，而 OPL 的伪标签辅助计算的使用进一步加强了类内特征的聚类和类间特征之间的分离。实验表明可以在域对齐的同时加强类别对齐，从而提供更好的性能。

**Table 3.** Accuracies (%) on VisDA-2017
**表 3.** VisDA-2017 数据集的精度(%)

| Classes | MCD [19] | ALDA [16] | CaCo [20] | SHOT [18] | STAR [21] | TVT* [22] | Baseline | SwinUDA |
|---|---|---|---|---|---|---|---|---|
| plane | 87.0 | 93.8 | 90.4 | 95.0 | 94.3 | 92.92 | **99.15** | 98.88 |
| bcycl | 60.9 | 74.1 | 80.7 | 84.0 | 88.5 | 85.58 | 80.37 | **89.50** |
| bus | 83.7 | 82.4 | 78.8 | 84.6 | 80.1 | 77.51 | 86.31 | **87.38** |
| car | 64.0 | 69.4 | 57.0 | **73.0** | 57.3 | 60.48 | 55.84 | 68.09 |
| horse | 88.9 | 90.6 | 88.9 | 91.6 | 93.1 | 93.60 | 98.19 | **98.64** |
| knife | 79.6 | 87.2 | 87.0 | 91.8 | 94.9 | 98.17 | 97.98 | **99.18** |
| mcycl | 84.7 | 89.0 | 81.3 | 85.9 | 80.7 | 89.35 | 94.93 | **95.68** |
| person | 76.9 | 67.6 | 79.4 | 78.4 | 80.3 | 76.40 | 70.93 | **81.08** |
| plant | 88.6 | 93.4 | 88.7 | **94.4** | 91.5 | 93.56 | 86.33 | 89.65 |
| sktbrd | 40.3 | 76.1 | 88.1 | 84.7 | 89.1 | 92.02 | 96.54 | **97.81** |
| train | 83.0 | 87.7 | 86.8 | 87.0 | 86.3 | **91.69** | 96.20 | **96.95** |
| truck | 25.8 | 22.2 | **63.9** | 42.2 | 58.2 | 55.73 | 44.30 | 58.70 |
| Avg | 71.9 | 77.8 | 80.9 | 82.7 | 82.9 | 83.92 | 83.92 | **88.46** |

为了证明模型具有广泛应用性，使用具有挑战性的 VisDA-2017 数据集，因为 152397 幅合成图像和 55388 幅真实图像之间存在显著的域偏移，示例图像如图 4 所示。评估了从合成图像到真实图像作为源域到目标域的方法。

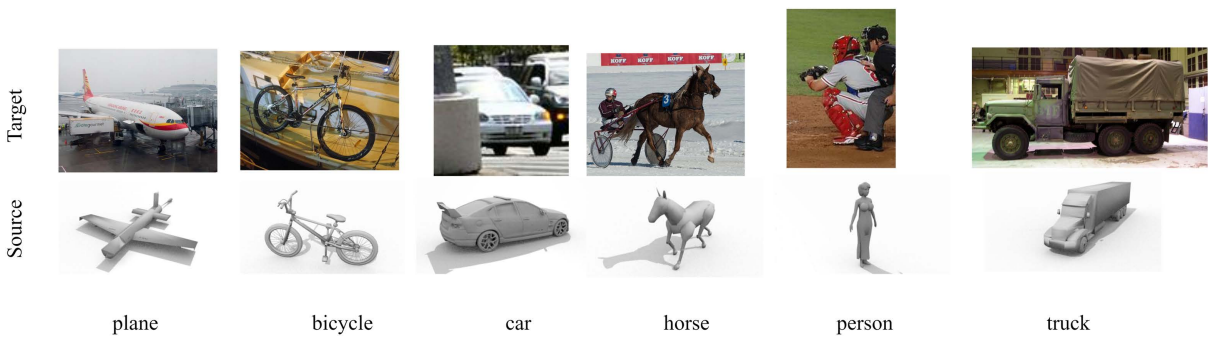



**Figure 4.** Example images of VisDA-2017 dataset
**图 4.** VisDA-2017 数据集的示例图像

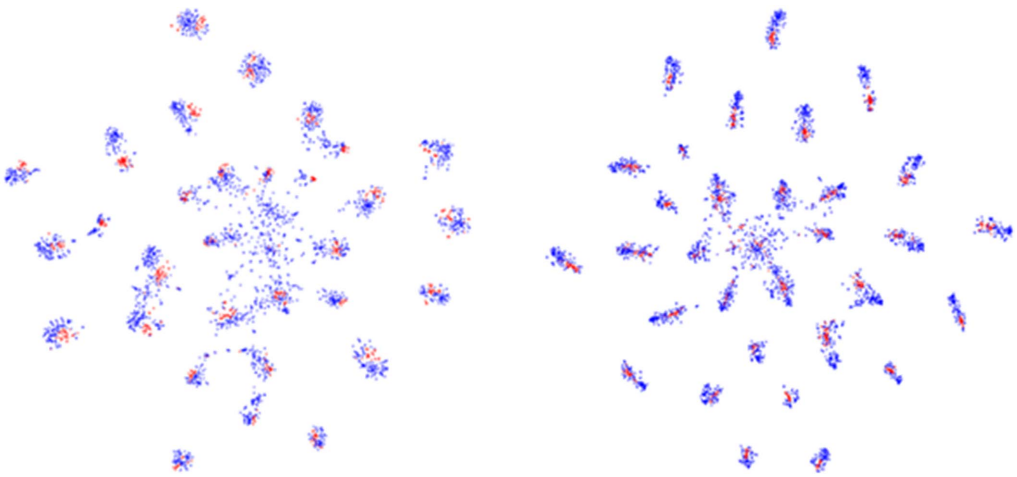
大规模 VisDA-2017 数据集的结果如表 3 所示，实验结果表明，与依赖伪标签的 ALDA 和 SHOT 相

比，提出的方法实现了更高的平均精度和进一步的改进。仔细观察结果，对于该数据集中的“人”，基线非常低，这表明基准模型在这一类别中的分类能力较差，这也说明了提出的方法的两个组件OPL和IML的有效性。相比之下，提出的方法对标签噪声具有一定的鲁棒性，这大大提高了实验结果。

为了了解IML和OPL两个组件的作用，进行的消融研究如表4所示。对于Baseline，IML持续提高分类精度，这表明捕获可转移和判别特征的重要性。引入OPL进一步提高了性能，证明了类内特征聚类的必要性。提出方法为真实的 VisDA-2017 数据集带来了大规模合成数据的最大改进。我们怀疑VisDA-2017中存在较大的域间隙是主要原因，因为简单地将两个域与较大的域偏移对齐会导致混乱的分布式特征空间。然而，IML可以解决这一挑战，它可以保留有区别的信息。同时OPL从类别级角度出发从特征空间进行约束，可以更好地聚类相似特征。可以观察到，IML和OPL是互补的，当移除任何一个组件时，性能都会下降。

**Table 4.** Ablation study of each module
**表 4.** 各模块的消融研究

| Method | Office-Home | Office-31 | VisDA-2017 | Avg |
|---|---|---|---|---|
| Baseline | 83.2 | 93.6 | 83.9 | 86.9 |
| +IML (wo OPL) | 85.9 | 94.4 | 84.0 | 88.1 |
| +OPL (wo IML) | 85.6 | 94.2 | 82.2 | 87.3 |
| SwinUDA (ours) | **87.2** | **94.6** | **88.5** | **90.1** |



(a) Baseline (b) SwinUDA

**Figure 5.** t-SNE of Pr → Cl task in the Office-31 dataset, where red and blue points indicate the source (synthetic rendering) and the target (real images) domain, respectively
**图 5.** Office-31 数据集中的 Pr → Cl 任务的 t-SNE，其中红色和蓝色点分别表示源(合成渲染)和目标(真实图像)域

## 5. 结论

本文提出了一种新的无监督域适应解决方案，即基于 Swin Transformer 的无监督域自适应(SwinUDA)。将Swin Transformer与简单的对抗性域自适应相结合进行域对齐，结果表明Swin Transformer具有强大的可迁移性。还考虑类级对齐，引入正交投影损失，并使用伪标签来计算该损失。这可以强制相同类特征的良好聚类和不同类特征的分离，导致来自相同类别的样本(无论来自源域或目标域)被拉近，而来自不同类别的样本被推开。为了保留目标域的内在结构，引入了互信息最大化损失来保留更多的目

标域信息，并进一步提高模型性能。所提出的方法在进行域对齐的同时保证了精确了类别对齐。大量实验表明，提出的方法优于现有方法。

## 基金项目

北京市自然科学基金(No. 8202013)；2022 年北京建筑大学研究生创新项目(NO. PG2022145)。

## 参考文献

[1] Ganin, Y. and Lempitsky, V. (2015) Unsupervised Domain Adaptation by Backpropagation. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 1180-1189.

[2] Long, M.S., Cao, Z.J., Wang, J.M., *et al*. (2018) Conditional Adversarial Domain Adaptation. *Advances in Neural Information Processing Systems* 31: *Annual Conference on Neural Information Processing Systems* 2018, Montréal, 3-8 December 2018, 31.

[3] Tzeng, E., Hoffman, J., Saenko, K., *et al*. (2017) Adversarial Discriminative Domain Adaptation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 7167-7176. https://doi.org/10.1109/CVPR.2017.316

[4] Cui, S., Wang, S., Zhuo, J., *et al*. (2020) Gradually Vanishing Bridge for Adversarial Domain Adaptation. *Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 12455-12464. https://doi.org/10.1109/CVPR42600.2020.01247

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. *Annual Conference on Neural Information Processing Systems* 2014, Montreal, 8-13 December 2014, 2672-2680.

[6] Zhang, Y., Tang, H., Jia, K., *et al*. (2019) Domain-Symmetric Networks for Adversarial Domain Adaptation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5031-5040. https://doi.org/10.1109/CVPR.2019.00517

[7] Jiang, X., Lao, Q., Matwin, S., *et al*. (2020) Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation. *International Conference on Machine Learning*, 13-18 July 2020, 4816-4827.

[8] Morerio, P., Volpi, R., Ragonesi, R., *et al*. (2020) Generative Pseudo-Label Refinement for Unsupervised Domain Adaptation. *IEEE Winter Conference on Applications of Computer Vision*, Snowmass, 1-5 March 2020, 3130-3139. https://doi.org/10.1109/WACV45572.2020.9093579

[9] Tang, H., Chen, K. and Jia, K. (2020) Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering. *Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 8725-8735. https://doi.org/10.1109/CVPR42600.2020.00875

[10] Saito, K., Ushiku, Y. and Harada, T. (2017) Asymmetric Tri-Training for Unsupervised Domain Adaptation. *International Conference on Machine Learning*, Sydney, 6-11 August 2017, 2988-2997.

[11] Xu, T., Chen, W., Wang, P., *et al*. (2021) CDTrans: Cross-Domain Transformer for Unsupervised Domain Adaptation.

[12] Shi, Y. and Sha, F. (2012) Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation.

[13] Saito, K., Kim, D., Sclaroff, S., *et al*. (2019) Semi-Supervised Domain Adaptation via Minimax Entropy. *International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 8050-8058. https://doi.org/10.1109/ICCV.2019.00814

[14] Ranasinghe, K., Naseer, M., Hayat, M., *et al*. (2021) Orthogonal Projection Loss. *International Conference on Computer Vision*, Montreal, 10-17 October 2021, 1233-12343. https://doi.org/10.1109/ICCV48922.2021.01211

[15] Peng, X., Usman, B., Kaushik, N., *et al*. (2017) VisDA: The Visual Domain Adaptation Challenges. https://arxiv.org/abs/1710.06924

[16] Chen, M., Zhao, S., Liu, H., *et al*. (2020) Adversarial-Learned Loss for Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 3521-3528. https://doi.org/10.1609/aaai.v34i04.5757

[17] Wang, X., Li, L., Ye, W., *et al*. (2019) Transferable Attention for Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 5345-5352. https://doi.org/10.1609/aaai.v33i01.33015345

[18] Liang, J., Hu, D. and Feng, J. (2020) Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. *International Conference on Machine Learning*, 13-18 July 2020, 6028-6039.

[19] Saito, K., Watanabe, K., Ushiku, Y., *et al*. (2018) Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. *Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3723-3732. https://doi.org/10.1109/CVPR.2018.00392

[20] Huang, J., Guan, D., Xiao, A., *et al*. (2022) Category Contrast for Unsupervised Domain Adaptation in Visual Tasks. *Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 1203-1214. https://doi.org/10.1109/CVPR52688.2022.00127

[21] Lu, Z., Yang, Y., Zhu, X., *et al*. (2020) Stochastic Classifiers for Unsupervised Domain Adaptation. *Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 9111-9120. https://doi.org/10.1109/CVPR42600.2020.00913

[22] Yang, J., Liu, J., Xu, N., *et al*. (2023) TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation. *IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 2-7 January 2023, 520-530. https://doi.org/10.1109/WACV56688.2023.00059

[23] Sun, T., Lu, C., Zhang, T., *et al*. (2022) Safe Self-Refinement for Transformer-based Domain Adaptation. *Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 7191-7200. https://doi.org/10.1109/CVPR52688.2022.00705

[24] Venkateswara, H., Eusebio, J., Chakraborty, S., *et al*. (2017) Deep Hashing Network for Unsupervised Domain Adaptation. *Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5018-5027. https://doi.org/10.1109/CVPR.2017.572

[25] Saenko, K., Kulis, B., Fritz, M., *et al*. (2010) Adapting Visual Category Models to New Domains. 11*th European Conference on Computer Vision*, Heraklion, 5-11 September 2010, 213-226. https://doi.org/10.1007/978-3-642-15561-1_16