个性化图像语义分割

Yu Zhang[1] Chang-Bin Zhang[1] Peng-Tao Jiang [1] Feng Mao[2] Ming-Ming Cheng[1*]
[1] 南开大学计算机学院 [2] 阿里巴巴

Abstract

在公共数据集上训练的语义分割模型近年来取得了很大的成功，然而，这些模型均没有考虑到个性化问题的分割，即使这在实践中是一个重要的问题，本文解决了个性化图像分割的问题。通过研究数据的个性化特征，在未标记的个性化图像上生成更准确的分割结果。为了开辟这一领域的未来研究，本文收集了一个大型数据集，其中包含了不同用户的个性化图像，称为 *PSS(个性化语义分割)*。本文还调查了与这个问题相关的一些最近的研究，并在提出的数据集上测试了它们的性能。此外，通过观察用户个性化图像之间的相关性，本文提出了一种基线方法，该方法能够在分割特定图像时加入图像间的上下文。大量的实验表明，该方法在数据集上优于现有的方法。代码和 *PSS* 数据集可通过***https://mmcheng.net/pss/*** 获取。

1. 简介

语义分割是计算机视觉领域研究的热点。该任务的目标是为给定图像的每个像素分配一个语义标签。与其他计算机视觉任务一样，深度学习以其强大的表示学习能力极大地提高了语义分割的能力 [3, 4, 12, 27, 29, 33, 52]。这些最先进的方法主要集中在 Pascal VOC [9]、ADE20K [55]、CityScapes [6] 等公开可用的数据集上，其中的图像被假设成是独立同分布的。然而，这种假设在现实世界中并不成立。例如，在移动端摄影中，用户可以通过拍照来记录自己的生活，形成个性化的影像集。一方面，个性化数据与公共数据集的分布不相同，导致在公共数据集上训练良好的分割模型存在泛化问题。另一方面，如 Fig. 1所示，来自同一用户的图像信息是相互关联的。它产生了潜在的研究价值，可以利用这一相互关联的性质优化分割结果。

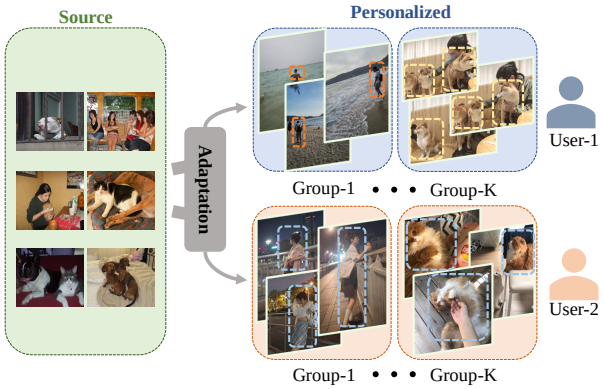


Figure 1. 提出的个性化图像数据集的样本。本文研究了如何利用用户图像的个性化属性从源数据进行自适应，可以观察到值得注意的一点，来自同一用户的图像是相关的 (类似的物体，场景等)。

本文讨论了个性化图像分割，这是一个在以前的工作中没有讨论过的问题。难点主要体现在以下两个方面:(i) 首先公共数据集与个人数据集的分布存在较大差距, 一个简单的方法是使用用户数据的额外注释来训练模型，这是非常困难的。因此，它迫切需要直接从未标记的个性化数据中学习。然而，没有可用的个性化数据集可供学习。(ii) 此外，来自同一用户的个性化图像通常具有一些个人特征。如何正确利用这些个性化特征在语义分割中仍然是一个未解决的问题。尽管存在上述困难，但在实践中对个性化图像分割的需求依然很大。例如，相机应用程序可能需要为用户的图像辅助摄影生成高质量的分割结果。为了解决这些问题，进一步开展个性化图像分割研究，本文提出了一个名为 PSS 的个性化数据集和基于该数据集的基线个性化分割方法。

*程明明是本文的通信作者

PSS 数据集包括 15 个用户的个性化图像，总共产生 10080 张图像。对于每个用户的个性化数据，随机选择大约 30% 的图像，并对其像素级语义进行注释以进行验证。为了更容易地从现有数据集调整到本文提出的个性化数据，本文考虑了 PASCAL VOC [9] 数据集中的 20 个常见目标类。据调查，PSS 数据集是第一个关注图像分割个性化问题的数据集，它使研究人员能够利用个性化特征来研究分割问题。

针对于个人特质的学习的挑战性，以往关于视觉任务个性化问题的研究 [15,19,35] 通常从个人数据中提取全局记忆来代表用户的偏好或个性，然后将提取的记忆作为特定下游任务的先验。然而，我们要从更广泛的角度考虑个性化，即每个人的个人特征都存在于每个人的图像中，没有必要将其提取出来作为一个全局代表。事实上，实验表明，在本文的场景中，为用户提取全局表示的方法是失败的。这种失败是合理的，因为在语义分割问题中，我们需要预测图像中每个像素的类别。虽然某些用户的图像有自己的特点，但他们的图像仍然是这些图像中的各种物体和场景。对所有像素的全局表示太不明确了。在学习个性化特征分割的同时，应该避免全局表示的模糊性。本文建议研究个性化图像之间的上下文联系，并在局部相关图像之间利用它们。具体来说，首先将一个人的图像分成几组，这样来自同一组的图像就可以共享相似的物体或背景。然后在每个组中，提取多个局部区域表示。对于该组中每个像素的预测，使用注意机制查询相关区域表征。

需要注意，在提出的个性化数据集中没有提供带标签的训练图像。本文通过域自适应来解决这个问题，从现有的标记数据集 (作为源) 到个性化图像 (作为目标)。关于无监督域自适应语义分割 (UDASS) 问题已有很多研究 [2,8,20,22,25,30,31,45]。虽然本文中的个性化图像是相互关联的，但目前的 UDASS 方法都将目标图像视为独立分布的。他们无法从个人数据中捕捉到个人特征。在本文的基线方法中，将一个组上下文模块合并到域适应框架中。它允许网络从现有的数据集适应个性化的图像，同时利用个性化图像中的个人特征。

本文的主要贡献主要有两点:

- 提出图像语义分割的个性化问题，并采集个性化图像数据集命名为 PSS，包含 15 个来自不同用户的数据。
- 本文选择了一些最近与这个问题相关的工作，并在本文的数据集上报告它们的性能。此外，提出了一种基线方法，通过学习局部区域表征来研究个人特征。该方法在提出的个性化数据集上实现了最先进的性能。

2. 相关工作

2.1. 个性化研究

个性化问题在许多计算机视觉和自然语言处理任务中都有讨论。 [32] 使用个性特征来增强机器翻译系统。 [15] 提出了一种个性化的食物图像分类器。 [35] 通过探索用户之前帖子的个性来预测社交媒体图片的标题和标签。 [19] 研究了基于用户偏好的图像增强。这些方法通常侧重于从现有数据中学习全局表示，在面对新数据时作为先验。本文探讨了个性化的图像语义分割，这是一个以前没有讨论过的问题。在我们提出的问题中，个性化可以从用户的全局特征和个性化图像的相关属性来研究。

2.2. 从相关数据学习

本文提出的个性化图像分割的一个关键挑战是从同一个人的相关图像中学习，即，提取互补语义，同时丢弃误导语义。联合分割 [11, 16, 23, 56] 和联合显著性检测 [10, 51] 的目的是挖掘分组图像中常见的语义对象，每组图像中包含相同类别的对象。Li 等人 [23] 提出了一种循环网络体系结构来探索常见的语义表示。Zhang 等人 [51] 利用共同的分类特征来发现目标的一般区域。这些方法通常学习每个组的组表示，作为从该组分割图像的先验。个性化的情况要复杂得多，因为个性化数据可能在不同的图像中包含不同的对象，需要分割所有的类而不是一个类。

2.3. 语义分割的域适应

个性化图像分割的目标是利用已有的数据集和模型预测未标记个性化图像的分割掩码。语义切分的无监督领域自适应是近年来研究较多的一个类似任务。本文将在其余部分称之为 UDASS。给定带标签的源数据集和未带标签的目标数据集，UDASS 致力于解决源数据集和目标数据集之间分布不匹配的问题，使模型能很好地从源到目标进行泛化。UDASS 的一项工作 [5,14,21,34,40–42,46] 使用基于对抗的方法来对齐源

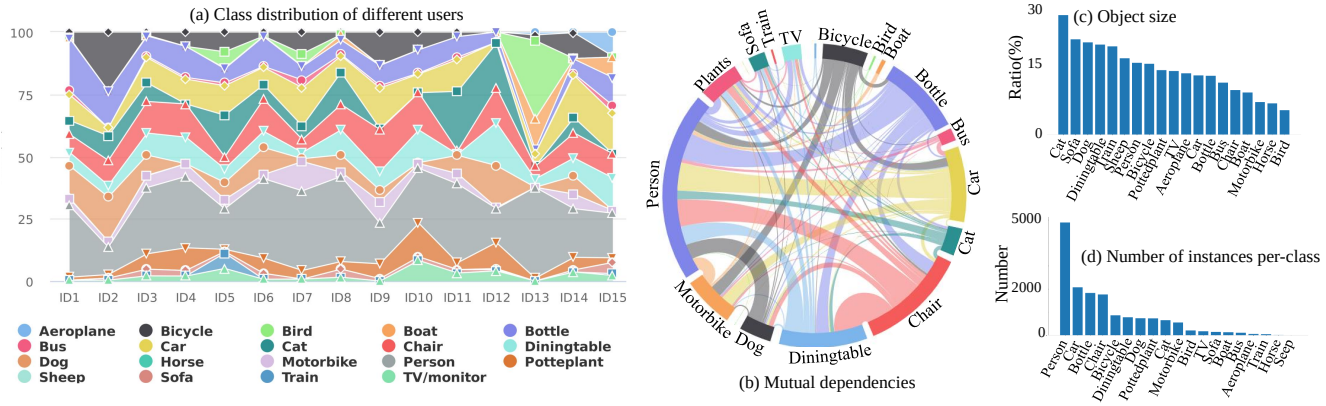


Figure 2. PSS 数据集的统计。(a) 每个用户的个性化数据中不同对象类的比例。(b) 数据集上不同类的相互依赖性。(c) 不同类的平均个数。(d) 每个类的实例数量。

和目标域的分布偏移。另一条工作线更侧重于学习策略: [17,18,24,26,43,49,50,57] 使用课程学习或自我训练策略来照看网络，以学习目标领域的良好语义。我们的问题和 UDASS 之间的主要区别是不独立地考虑目标域的图像。相反，认为来自同一个人的图像是相关的部分。类似的图像可以为其他人提供有用的信息。此外，目前的 UDASS 方法主要集中在城市场景数据集，如 Cityscapes [6]、GTA5 [37] 和 SYNTHIA [38]，在这些数据集中进行合成图像和真实图像之间的域适配。在提出的数据集中，我们重点关注普通对象的个性化图像。数据集提供了更加多样化和现实的个性化场景，也可以评估域适应方法的有效性。

2.4. 语义分割数据集

与其他计算机视觉任务一样，数据集在图像分割研究中起着关键作用。最近的数据集极大地支持了基于深度学习的分割框架 [3, 12, 52]。PASCAL VOC [9] 和 COCO [28] 是针对常见物体图像的数据集。ADE20K [55] 也关注普通对象，但使用更细粒度的类标签，如对象部分。CityScapes [6], GTA5 [37], SYNTHIA [38] 和 Synscapes [44] 是城市场景的数据集。虽然最近提出了许多数据集，但这些数据集都没有考虑到分割中的个性化问题。在本文中，我们收集了不同用户的 PSS 数据集。本文的数据集中于具有不同用户个性化特征的常见物体的图像，是一个很好的个性化分割研究的开始，也可以提供一个很好的基准用于其他分割任务，如域适应。

3. 提出的数据集

3.1. 数据采集

为了模拟真实世界的个性化数据分布，我们直接从不同的志愿者那里收集数据集。每位志愿者被要求将手机或相机中的图像导出，形成个人数据。为了保护志愿者的隐私，要求志愿者浏览图片，并过滤掉不愿意公开的图片。本文的数据集中于 PASCAL VOC [9] 中的 20 个类。最终，我们得到了由 15 个用户的个性化数据组成的 10080 幅图像的大规模数据集。每个个性化的数据可能具有与其他用户不同的数据分布，并且可能具有一些对语义分割有用的高级/低级统计信息。

3.2. 数据标注

我们请了几位训练有素的专家对收集到的个性化数据进行注释，数据集提供了图像级和像素级注释。

图像级别标注. 与 [9] 一致，数据集中的所有图像都用出现的对象的类标签进行标记。图像级注释可用于数据分析。在 Fig. 2(a) 和 Fig. 2(b) 中展示了每个用户的对象类分布以及不同类之间的相互依赖关系。

像素级标注. 个性化图像分割的难点是在未标记的个性化图像上生成分割掩码。为了在本文的数据集上进行模型评估，为每个用户数据的约 30% 提供像素级注释。对于每一个像素标注的图像，目标区域属于 20 个类。在 PASCAL VOC [9] 中，它们被标记了特定

的值，表示图像中每个像素的类别的逐像素掩模。在 Fig. 2(c) 和 Fig. 2(d) 中显示了不同类的平均大小和实例数量。

3.3. 数据集特点

个性化数据.

本文数据集最重要的特征是个性化，这自然会导致用户内部一致性，即某些用户的数据有其特点，可能是一致的不同图像，这可以用来促进学习。另一方面，不同用户的图像在低水平 (如光照条件、图像质量) 和高水平 (如图像内容、背景) 属性上也有所不同。不同用户之间数据分布的差异需要细分模型来适应特定的用户数据。更多关于用户内部一致性和用户间分布差距的细节可以在补充材料中找到。

真实数据.

我们的个性化数据集非常接近现实场景。现实存在于两个层面。首先，本文的数据集是直接从不同的用户那里收集的。这些图像真实地反映了他们在日常生活中所关心和拍摄的东西，这意味着数据集的结果可以反映出不同练习方法的有效性。如补充资料中所示的数据示例: 一些用户拥有更多关于日常生活的食物或宠物图片，而另一些用户拥有更多关于美丽风景的图片。这表明了个性化细分的重要性。其次，本文数据集中的对象类是长尾分布的，如 Fig. 2(d) 所示。有些物体更容易被拍摄，而另一些则不会。在大多数图片中可能有“人”，而只有少数“船”的实例。如何解决类分布不均的问题是一个值得探索的有趣方向。

4. 提出的方法

在这一节中，将介绍本文的个性化图像分割方法。

综述. 考虑图像 $\{I_s \subset \mathbb{R}^{3\times H\times W}\}$ 及其 c 类分割标签 $\{L_s \subset \mathbb{R}^{C\times H\times W}\}$ 的源数据，未标记的个性化数据 $\{I_p \subset \mathbb{R}^{3\times H\times W}\}$。本文方法的关键思想是通过使用来自同一用户的其他图像的上下文来利用个性化图像 $\{I_p\}$ 之间的相关性。在 Fig. 3中展示了方法的架构。本文的个性化图像分割框架有两个主要步骤: 域自适应步骤和随后的伪标签细化步骤。在第一步中，用一个基于对抗性的领域自适应框架从源数据适应到个性化数据。在训练过程中，结合本文提出的组区域上下文模块，利用个性化数据中的图像间上下文。在第二步中，选择个性化数据中容易识别的图像作为带有熵的伪标签，利用伪标签作为易分割图像的真值，指导分割网络。

4.1. 基于对抗的域自适应

首先介绍在第一步中使用的基于对抗的域自适应技术。S 表示一个分割网络，取 I_s 作为输入，输出一个软预测图 $P_s = S(I_s) \in \mathbb{R}^{C\times H\times W}$，其中每个值 $P_s^{(c,h,w)}$ 表示像素 $I_s^{(h,w)}$ 属于类 c 的概率，参考真值 Y_s，交叉熵损失定义为:

$$\mathcal{L}_{seg} = -\sum_{h,w}\sum_{c} Y_s^{c,h,w} log(P_s^{(c,h,w)}) \tag{1}$$

优化交叉熵来训练分割网络。除分割损失外，采用对抗性训练对齐源数据 $\{I_s\}$ 与个性化数据 $\{I_p\}$ 之间的分布差异。给定源图像和个性化图像的分割预测 P_s 和 P_p，用下式计算它们的熵图

$$E_s^{h,w} = \sum_{c} -P_s^{c,h,w} log(P_s^{c,h,w}). \tag{2}$$

一个用于预测 E_s 和 E_p 域标签的判别器 D。通过训练分割网络 S 欺骗 D，可以缩小来源预测和个性化数据之间的分布差距。对抗性损失的公式为:

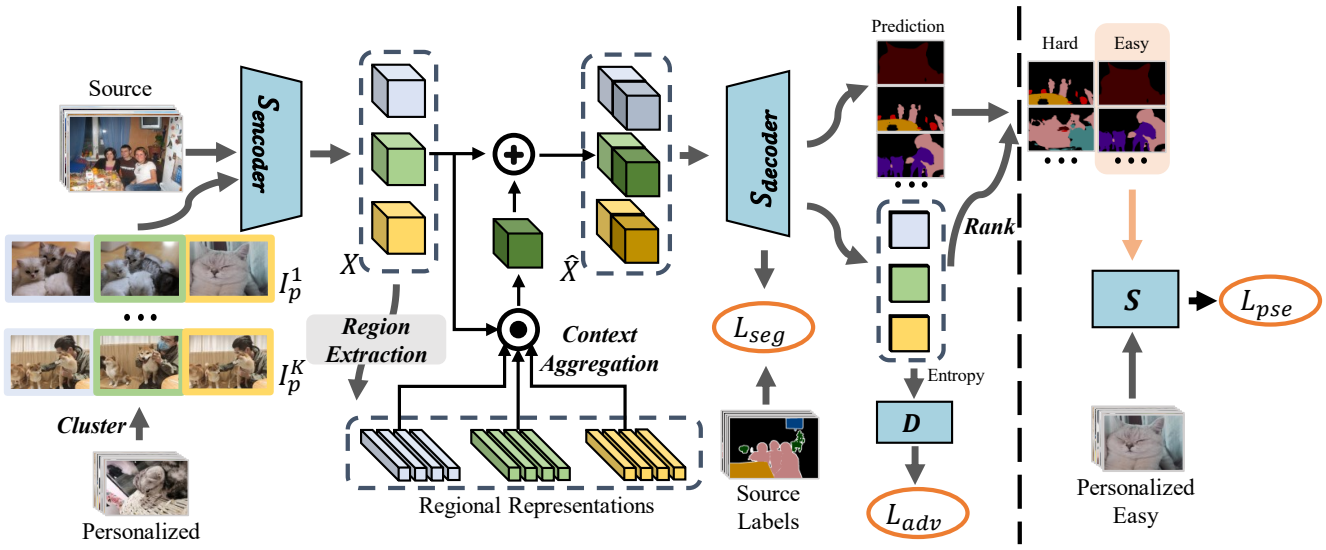
$$\mathcal{L}_{adv}(I_s, I_p) = -\sum_{h,w} log(1 - D(E_s^{h,w})) + log(D(E_p^{h,w})). \tag{3}$$

这种对抗性范式可以调整源数据和个性化数据之间的分布不匹配。但是，它单独获取个性化数据中的每一幅图像, 因此没有考虑到 $\{I_p\}$ 内部的相关性。为此，本文提出了一个组区域上下文模块，用于获取个性化数据的图像间上下文。

4.2. 组上下文模块

本文设计了一个简单的组上下文模块，利用提出的个性化数据集的相关属性，首先将每个用户的个性化数据聚到多个组中。每个组包含具有相似语义的图像。在每一组中，提取所有图像的软区域上下文表示。在分割过程中，所有基于软区域的上下文表示都被推断出来以帮助训练。

对于用户的个性化数据 $\{I_p\}$，将其输入到 ImageNet [13] 上预训练的 ResNet-50 [7] 中，并在最后一个完全连接层之前得到表示 $\{F_p \in \mathbb{R}^{2048}\}$，然后在 $\{F_p\}$ 上采用 K-means 聚类算法，得到 K 组图像在 $\{\{I_p^1\}, \{I_p^2\}, \cdots, \{I_p^K\}\}$。将分割网络考虑为编码器 $S_{encoder}$ 和解码器 $S_{decoder}$ 的组合。编码器从 k 组中



(a) Domain Adaptation with Group Context (b) Pseudo Label Refine

Figure 3. 本文的个性化图像分割方法，本文的模型包含两个步骤。第一步是域自适应步骤，如 (a) 所示。在第二步中，进一步像 (b) 一样添加一个伪标签损失 L_{pse}。在 (a) 中，首先将个性化数据聚类到 K 组中。然后在每一组中，用区域组上下文增强图像表征 X，得到 $\hat{X}$。为简单起见，仅为每个组显示三幅图像，并且仅为标记为绿色的图像显示组区域上下文聚合过程。

选取 I_p 作为输入，输出中间表示 $X = S_{encoder}(I_p) \in \mathbb{R}^{CH\times W\times H}$，其中 CH 和 W, H 分别表示 X 的通道大小和空间大小。本文的组上下文模块利用组上下文 $\mathrm{d}F_{group}$ 学习一个增强的表示 $\hat{X} = F_{group}(X) \in \mathbb{R}^{CH\times W\times H}$。在组上下文模块中有两个步骤: 区域上下文提取和组区域上下文聚合。

区域背景提取.

受 [48] 的启发，本文将划分了 I_p 划分为 C 个软对象区域。C 是对象类的数量。使用 $\mathrm{aux}P_p \in \mathbb{R}^{C\times W\times H}$ 输出分割网络。计算每个软区域的表示为

$$f_c = \sum_i r_{ci} X_i, \tag{4}$$

其中 i 表示空间位置，X_i 表示像素 i，r_{ci} 是 $P_{pi} \in \mathbb{R}^C$ 的 softmax 归一化计算的像素权重为 $r_{ci} = softmax(P_{pi})_c$。对于一组 N 个图像，可以提取 $N\times C$ 个区域表示这个组。

组区域上下文聚合 Given a group's region representations $\{f_{i,j}|i \in [1,C], j \in [1,N]\}$, we compute group context representation for each pixel in X by weighted aggregation of group regions: 给出一组区域代表 $\{f_{i,j}|i \in [1,C], j \in [1,N]\}$，用区域权重聚合的方式为 X 中的每个像素计算组上下文表示。

$$c_{h,w} = \rho(\sum_{i,j} w_{(i,j),(h,w)} \sigma(f_{i,j})). \tag{5}$$

这里 ρ 和 σ 是两个线性变换函数，权重 $w_{(\tilde{i},\tilde{j}),(h,w)}$ 通过测量像素 $X_{h,w}$ 和区域表示 $f_{\tilde{i},\tilde{j}}$ 之间的关系来计算的

$$w_{(\tilde{i},\tilde{j}),(h,w)} = \frac{e^{s(X_{h,w}, f_{\tilde{i},\tilde{j}})}}{\sum_{i\in[1,K], j\in[1,N]} e^{s(X_{h,w}, f_{i,j})}}, \tag{6}$$

式中 $s(X_{h,w}, f_{i,j})$ 是形如 $s(X_{h,w}, f_{i,j}) = \phi(X_{h,w})^T \varphi(f_{i,j})$ 的隶属度函数，ϕ 和 φ 是由一个全连接层构成的转换函数。

在获得组上下文后，可以将像素表示增强为:

$$\hat{X}_{h,w} = \psi([X_{h,w}, c_{h,w}]). \tag{7}$$

$[*,*]$ 表示连接，ψ 是线性转换。结果表示 $\hat{X}$ 将被输入解码器并输出预测图:$\hat{P}_p = S_{decoder}(\hat{X})$。对于 X 中的每个像素，组区域上下文增强模块将与组上下文在同一组中的相似区域的表示集合起来，为分割网络提供额外的信息。

4.3. 使用伪标签进行优化

除了第一步的域自适应外，目前用于语义分割的域自适应方法 [34, 53] 通常采用伪标签来进一步细化网络。同时，在本文的方法中采用这种训练范式来预

Method	Backbone	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
No-DA		45.39	51.99	48.95	47.60	58.03	48.15	56.86	62.45	48.23	45.14	62.37	51.68	48.56	48.13	41.57	51.01
AdaptSeg [40]		46.87	52.16	50.06	48.51	59.78	51.39	57.12	63.41	50.99	46.15	60.68	52.84	50.32	50.69	43.08	52.27
MaxSquare [5]		48.28	52.50	50.61	50.54	61.39	54.60	59.36	63.43	50.67	46.49	62.94	52.68	49.65	48.99	46.00	53.20
FDA [47]	ResNet-50	50.12	53.70	53.22	50.76	60.29	55.01	58.18	65.89	53.28	46.49	62.09	56.10	48.93	51.38	47.03	54.16
ADVENT [42]		53.39	57.33	52.42	52.51	64.63	55.04	60.61	61.69	55.34	49.18	66.05	57.83	56.04	54.38	52.34	56.59
MRNet [54]		**54.05**	58.62	54.29	53.17	61.72	57.24	62.20	66.46	56.75	50.27	66.76	54.20	53.87	54.38	51.38	57.02
OURS-S1		52.90	59.12	54.74	55.82	64.97	**60.38**	61.78	68.12	56.99	51.21	69.42	60.44	57.05	54.41	54.51	58.79
OURS-S2		53.28	**60.39**	**54.81**	**56.02**	**66.87**	60.11	**63.77**	**69.09**	**57.44**	**52.66**	**70.42**	**60.77**	**58.50**	**56.84**	**54.85**	**59.72**
No-DA		33.68	33.56	35.50	35.49	39.52	37.55	36.23	47.95	34.35	32.86	50.95	41.48	39.24	30.90	34.51	37.58
AdaptSeg [40]		32.70	37.65	37.16	33.54	40.55	41.11	43.17	52.12	36.95	31.83	49.04	40.97	33.54	31.49	34.06	38.39
MaxSquare [5]		36.17	32.99	38.81	37.36	42.64	42.03	49.88	50.06	37.99	35.93	51.33	41.98	36.27	36.35	37.13	40.46
FDA [47]	VGG-16	34.61	36.75	35.53	36.60	38.36	40.07	45.21	52.57	37.79	35.01	49.59	41.93	33.72	35.01	36.27	39.27
ADVENT [42]		39.89	44.39	39.88	40.01	49.89	44.24	47.99	54.59	43.84	38.29	53.00	43.07	42.83	40.02	41.36	44.22
MRNet [54]		34.40	41.18	36.67	32.18	44.63	38.12	41.99	46.78	39.51	36.54	39.39	44.17	35.93	37.17	38.35	39.13
OURS-S1		41.87	45.73	43.14	**44.04**	52.44	47.45	**52.32**	56.92	45.61	**42.67**	54.94	**48.38**	44.24	41.67	45.98	47.16
OURS-S2		**43.24**	**47.89**	**44.67**	44.00	**53.27**	**50.68**	52.18	**57.86**	**46.84**	42.34	**56.56**	46.28	**47.02**	**42.98**	**47.01**	**48.19**

Table 1. 使用 ResNet-50 和 VGG-16 对不同方法的 FIoU 结果进行比较。列号表示 15 个用户 id。“Mean” 列表示总体平均性能 id。

Method	Backbone	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
No-DA		28.05	29.18	30.78	33.05	42.52	31.31	35.85	28.63	39.60	36.99	33.15	38.51	29.78	32.75	31.85	33.47
AdaptSeg [40]		31.69	28.87	30.50	35.09	45.83	32.55	36.70	33.83	36.43	36.49	34.09	41.23	31.02	35.52	34.40	34.95
MaxSquare [5]		28.72	28.91	31.81	36.45	40.09	33.94	38.85	31.21	35.85	32.23	28.58	34.16	33.58	30.35	34.78	33.30
FDA [47]	ResNet-50	31.94	31.16	32.39	36.11	45.35	35.76	37.46	30.93	42.91	**43.28**	**37.51**	38.09	29.31	**37.25**	35.76	36.35
ADVENT [42]		36.04	34.04	36.98	39.98	43.76	40.52	41.59	29.69	36.26	39.19	33.46	39.05	38.17	33.43	37.44	37.31
MRNet [54]		**38.27**	**35.02**	36.98	36.54	43.99	**40.90**	40.22	36.26	32.35	33.10	36.26	31.78	37.77	35.89	32.24	36.51
OURS-S1		36.61	34.43	31.88	40.36	44.25	33.64	38.14	32.25	39.87	38.69	37.20	42.44	**39.60**	30.37	**42.18**	37.46
OURS-S2		33.85	33.38	**38.40**	**41.36**	**46.73**	37.58	**44.19**	**36.87**	**44.66**	42.03	37.42	**43.71**	35.12	34.18	37.89	**39.16**
No-DA		15.78	17.19	17.80	21.41	21.54	18.35	19.07	15.40	21.67	22.80	18.55	21.06	21.66	18.96	22.95	19.61
AdaptSeg [40]		16.59	19.04	18.96	23.81	21.43	23.12	25.47	16.26	23.33	22.60	19.08	20.20	22.12	20.10	24.30	21.09
MaxSquare [5]		18.46	18.19	18.46	22.29	26.01	23.88	25.36	17.07	25.01	25.37	19.99	20.56	24.52	20.82	25.90	22.13
FDA [47]	VGG-16	17.17	17.82	20.07	24.44	23.82	25.69	25.22	15.56	23.31	25.53	21.14	20.68	20.82	19.26	24.65	21.68
ADVENT [42]		27.32	21.95	**25.21**	25.46	35.50	23.75	28.18	**25.03**	**32.47**	29.73	24.76	25.00	26.74	24.76	26.31	26.81
MRNet [54]		20.30	16.46	20.92	27.62	29.70	25.46	27.74	22.30	30.64	26.46	17.64	27.12	23.43	23.86	26.08	24.38
OURS-S1		24.40	**24.93**	20.31	31.01	35.54	30.67	28.81	20.68	27.45	**32.06**	27.63	**31.44**	27.68	23.87	**33.53**	28.00
OURS-S2		**25.53**	24.40	22.62	**33.55**	**35.73**	**31.86**	**32.14**	21.84	28.62	30.57	**30.26**	24.97	**28.82**	**25.25**	31.91	**28.54**

Table 2. 采用 ResNet-50 和 VGG-16 为骨干网络，对不同方法的 MIoU 结果进行分析。列号表示 15 个用户 id。“Mean” 列表示总体平均性能 id。

测个性化的图像。前面介绍的熵图 E_p 是图像 I_p 分割网络不确定性的一个指标。低不确定性的预测通常意味着输入图像简单，结果可靠性高。因此，选择具有低熵值的预测作为伪标签。请注意，与 VOC 和 cityscape 等数据集相比，每个人的个性化数据相对较小，单独这些数据集没有足够的数据来训练分割网络。因此，与 [34] 不同的是，本文使用额外的分段损失 $\mathcal{L}_{pse}$ 将伪标签添加到网络中，而不是使用伪标签替换源数据集。

5. 实验

5.1. 数据集和评估指标

本文收集的个性化数据集以拥有与 PASCAL VOC [9] 相同的类。因此在训练过程中，使用增强的 VOC 训练集作为源数据集，该数据集包含 10582 张已标记的图像，包含 20 类目标。定量评价采用均值交叉 (MIoU)。注意，个性化数据通常是长尾分布的，这意味着类是非常不平衡的。MIoU 可能会因为这种不平衡而失真。因此，我们进一步使用另一个度量，称为前景交叉的联合 (FIoU)。FIoU 反映的是图像而不是类的平均欠条。具体来说，我们首先计算前景图像 i 的 IoU_i，然后计算所有图像 $\sum_i^N IoU_i$ 的平均 IoU。

Methods	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
None	39.89	44.39	39.88	40.01	49.89	44.24	47.99	54.59	43.84	38.29	53.00	43.07	42.83	40.02	41.36	44.22
Global	38.89	42.13	42.13	38.96	50.88	45.09	51.62	55.67	44.23	38.51	49.80	45.70	43.20	39.84	41.71	44.56
OURS	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16

Table 3. 组上下文模块的消融。” None ” 和” Global ” 分别表示没有上下文和全局上下文。

Groups	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
1	42.39	46.11	42.95	43.79	52.45	46.15	51.23	55.88	45.57	42.51	55.30	47.89	43.66	39.54	44.12	46.64
10	42.49	45.52	42.75	44.07	52.64	46.54	50.85	56.52	45.81	42.25	54.75	46.86	44.62	41.18	45.36	46.81
80	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16
200	42.11	45.66	43.14	43.33	52.08	45.85	52.05	56.51	45.78	42.13	53.75	46.73	43.80	41.59	45.35	46.66

Table 4. 不同数量组的消融研究，列表示不同的用户 id、FIoU 记录。

5.2. 实现细节

我们使用 ImageNet [7] 预训练的 ResNet50 [13] 作为分割网络的骨干网络。一个 PSP 模块 [52] 被配备到分割网络如 [42]。自适应训练的输入是源图像和标签、分组目标图像。为了简化训练和节省计算，我们没有使用组中的所有图像区域来构建组上下文。相反，我们将每个批处理限定在同一组中，然后使用每个批处理中的图像来计算它。图像增强采用随机裁剪。所有的输入在训练期间都被调整为 320×320。在伪标签细化步骤中，选择率 $r = 0.5$ 用于选择可靠的预测。伪标签像素被设置为 255。为了简化训练过程并节省 GPU 内存，本文没有在一次迭代中处理整个组。相反，只需确保所有操作的图像都来自同一组。本文将所有实验的批量设置为 8。使用 SGD 优化器 [1]，学习率 2.5×10^{-4}，动量和权重衰减设置为 0.9 和 10^{-4}。代码是用 PyTorch [36] 库实现的。

5.3. 性能比较

本文记录了一些选定的域自适应方法在数据集上的性能，包括 AdaptSeg [40], MaxSquare [5], FDA [47], ADVENT [42] 和 MRNet [54]。这些方法都是对目标图像进行单独处理，而不考虑个性化图像的相关性。所有模型以 VOC [9] 为源数据库，以个性化数据为目标进行训练。MRNet [54] 等方法仅在步骤 2 中使用目标伪标签对分割网络进行监督，由于我们的个性化数据数量相对较少，导致性能较差。所以增加了对这种方法的 VOC [9] 标签的额外监督。在个性化数据集的注释验证分割上测试结果。我们将 FIoU 和 MIoU 的结果分别列于 Tab. 1和 Tab. 2。我们将不需要伪标签细化的方法和完整模型分别表示为 *OURS-S1* 和 *OURS-S2*。

总的来说，以 ResNet50 [13] 为骨干网络，*OURS-S1* 获得 37.46 MIoU 和 58.79 FIoU。与基线方法 ADVENT 相比，性能分别提高了 0.15 和 2.20，表明了本文的组上下文模块的有效性。值得注意的是，MIoU 0.15 的改进相对 FIoU 来说是微小的。我们推测这是由个性化数据的长尾特性造成的。由于组上下文模块合并了其他图像的上下文以帮助学习，它往往在有许多图像的类上表现得更好，但可能会降低稀有类的结果。当评估 MIoU 时，结果可能会受到这些稀有类的影响。我们在补充材料中为不同的用户提供了 class IoU 结果。通过使用伪标签，*OURS-S2* 分别获得 39.16 MIoU 和 59.72 FIoU，性能分别提高了 1.7 和 0.93。我们在 Fig. 4中显示了一些预测的掩模。

6. 讨论

6.1. 组上下文的有效性

在本节中，我们通过比较两个基线:*None* 和 *Global* 来研究我们提出的群组上下文模块的有效性。*None* 表示直接从编码器中使用特性 X 而不需要上下文。*Global* 表示使用全局组上下文来增强对象联合分割方法 [23] 中的表示。本实验骨干网络采用 VGG-16 [39] 进行。如 Tab. 3所示，*None* 基线平均达到 44.22 FIoU。*Global* 稍微提高了 0.34 的性能，这表明在本文的情况下，全局组表示不够有效。*OURS* 提高了 2.94 的性能，这表明了所提出的组上下文模块的有效性。

6.2. 个性化训练的价值

在本节中，我们合并所有用户的所有图像，形成一个大的数据集 *MixAll*，然后从 *MixAll* 中随机抽样 672

Methods	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
MixSample	40.92	43.18	40.73	40.55	49.84	46.18	50.42	57.22	42.92	39.39	54.26	45.83	43.67	38.74	44.90	45.25
MixAll	42.54	45.11	41.73	39.87	49.58	43.74	52.64	56.88	43.82	38.55	55.26	47.18	42.33	38.97	45.16	45.56
Personal	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16

Table 5. 混合图像集实验。“MixAll”表示混合所有用户的图像，“MixSample”从“MixAll”中抽取 1/15 的样本，使每个个性化数据具有相似的大小。

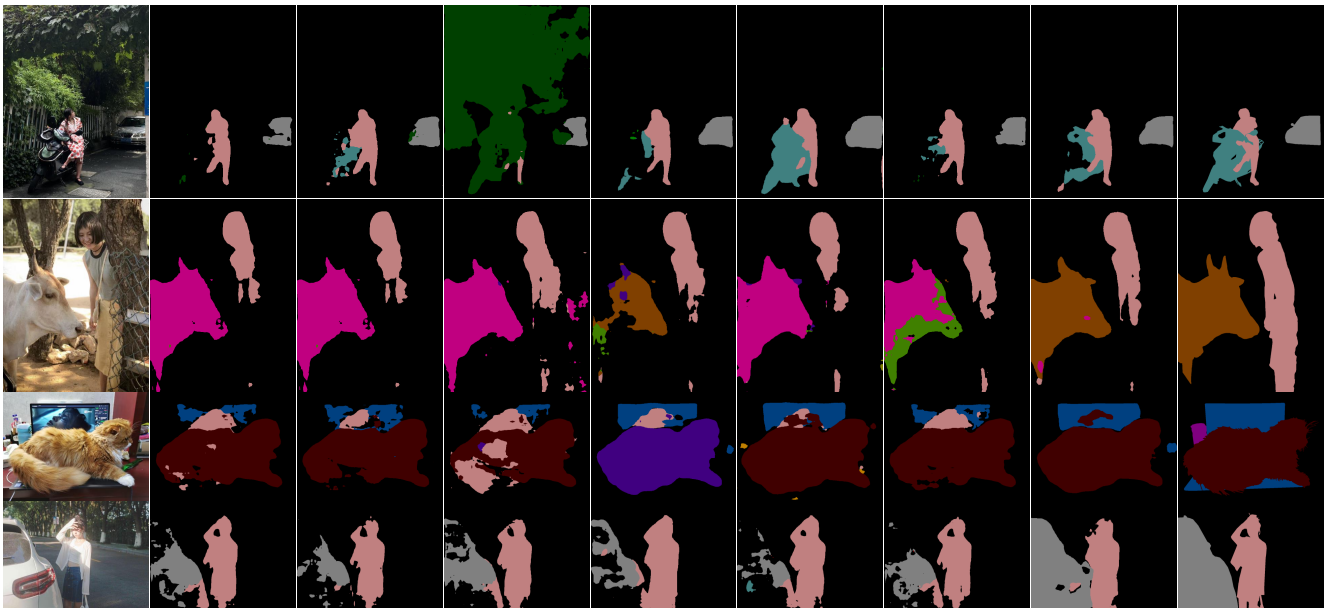


Image No-DA AdaptSeg [40] FDA [47] ADVENT [42] MRNet [54] MaxSquare [5] Ours GT

Figure 4. 不同方法的质量比较.

个图像作为子集。在这些图像集上训练模型，并对模型对不同用户数据进行评估，结果如 Tab. 5所示。利用大约 15 倍的目标图像，*MixAll* 达到 45.56 FIoU，低于 47.16 的 *Personal*，从相应的个性化数据进行培训。结果显示了个性化数据学习的价值。

6.3. 组的数量

在本节中，我们将每个用户的个性化数据聚类成不同数量的组，并研究组的数量如何影响推理时的分割性能。如 Tab. 4所示。不同的行表示不同数量的组。在组数为 1 的情况下，某些用户的所有图像都被视为在一个组中。在计算一组区域上下文时，可能会考虑不相关的图像，造成网络混乱。当组数为 200 时，每个组中的图片数量太少，无法为组上下文模块提供足够的上下文。在组数为 80 的情况下，平均 FIoU 为 47.16，优于其他数。但是，我们仍然可以注意到不同的用户使用不同的组号可以得到最好的结果。推测这是由于不同用户之间的分布差距造成的。结果表明，不同的用户需要不同数量的组。在未来，我们将研究更灵活的方法来聚类个性化图像而不是使用固定的组数。

7. 结论

本文主要研究图像语义分割中的个性化问题。首先收集了一个大型的个性化图像数据集 PSS，其中包含 15 个用户的数据。本文的数据集可以作为研究细分中的个性化问题的良好开端。个性化图像分割问题的挑战有两个方面。一是如何从不同用户的未标记数据中学习；另一个问题是如何利用特定用户数据中的个性化特征。利用个性化图像的相关特性，提出了一种采用图像间上下文的基线分割方法。在未来的工作中，我们将探索更复杂的方法来从未标记的数据中学习，还将研究如何在稀有类中提高组上下文模块的性能。

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, pages 177–186. Springer, 2010.

[2] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1900–1909, 2019.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Int. Conf. Comput. Vis.*, pages 2090–2099, 2019.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[8] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 982–991, 2019.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[10] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[11] Junwei Han, Rong Quan, Dingwen Zhang, and Feiping Nie. Robust object co-segmentation using background prior. *IEEE Trans. Image Process.*, 27(4):1639–1651, 2017.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[15] Shota Horiguchi, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. Personalized classifier for food image recognition. *IEEE Trans. Multimedia*, 20(10):2836–2848, 2018.

[16] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[17] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 705–722, 2020.

[18] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2020.

[19] Han-Ul Kim, Young Jun Koh, and Chang-Su Kim. Pienet: Personalized image enhancement. In *Eur. Conf. Comput. Vis.*, 2020.

[20] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12975–12984, 2020.

[21] Minsu Kim, Sunghun Joung, Seungryong Kim, JungIn Park, Ig-Jae Kim, and Kwanghoon Sohn. Cross-domain grouping and alignment for domain adaptive semantic segmentation. In *The National Conference on Artificial Intelligence (AAAI)*, 2021.

[22] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. In *The National Conference on Artificial Intelligence (AAAI)*, 2021.

[23] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *Int. Conf. Comput. Vis.*, pages 8519–8528, 2019.

[24] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 440–456, 2020.

[25] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[26] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Int. Conf. Comput. Vis.*, pages 6758–6767, 2019.

[27] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1925–1934, 2017.

[28] Tsungyi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015.

[30] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 6778–6787, 2019.

[31] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[32] Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. Motivating personality-aware machine translation. In *Conf. Empir. Meth. Natur. Lang. Process.*, 2015.

[33] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[34] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3764–3773, 2020.

[35] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6432–6440, 2017.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019.

[37] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Eur. Conf. Comput. Vis.*, pages 102–118, 2016.

[38] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3234–3243, 2016.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2015.

[40] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7472–7481, 2018.

[41] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Int. Conf. Comput. Vis.*, pages 1456–1465, 2019.

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2517–2526, 2019.

[43] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12635–12644, 2020.

[44] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018.

[45] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Eur. Conf. Comput. Vis.*, 2018.

[46] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[47] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4085–4095, 2020.

[48] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[49] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[50] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Int. Conf. Comput. Vis.*, pages 2020–2030, 2017.

[51] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, 2020.

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[53] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. In *In-*

ternational Joint Conference on Artificial Intelligence (IJCAI), 2019.

[54] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *Int. J. Comput. Vis.*, 2020.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[56] Chenyang Zhu, Kai Xu, Siddhartha Chaudhuri, Li Yi, Leonidas J. Guibas, and Hao Zhang. Adacoseg: Adaptive shape co-segmentation with group consistency loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[57] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Eur. Conf. Comput. Vis.*, pages 289–305, 2018.