基于扩散感知的文本图像对齐

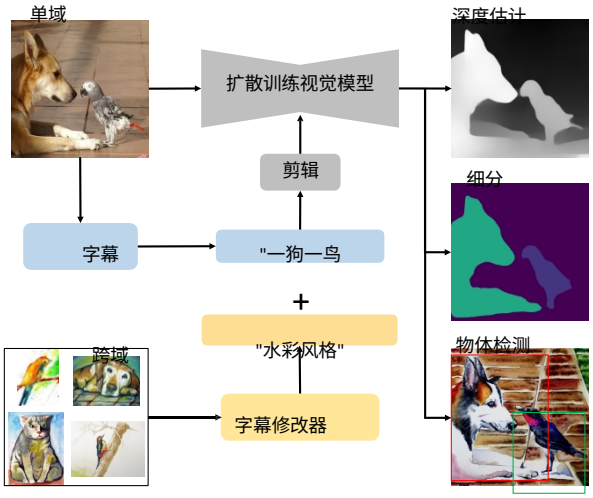
Neehar Kondapaneni[1] * Markus Marks[1*] Manuel Knott[1,2*]
Rogerio Guimaraes[1] Pietro Perona[1]

[1]加州理工学院[2] 苏黎世联邦理工学院
、瑞士数据科学中心、Empa

摘要

扩散模型是一种生成模型，具有令人印象深刻的文本到图像的合成能力，并为经典的机器学习任务带来了新一轮的创造性方法。然而，利用这些生成模型的感知知识来完成视觉任务的最佳方法仍是一个悬而未决的问题。具体来说，目前还不清楚在视觉任务中应用扩散背骨时如何使用提示界面。我们发现，自动生成的字幕可以改善文本与图像的对齐，并显著增强模型的交叉注意图，从而提高感知性能。我们的方法超越了目前在 ADE20K 上基于扩散的语义分割方面的最先进水平（SOTA），以及目前在 NYUv2 上深度估计方面的总体 SOTA 水平。此外，我们的方法还可以推广到跨域集

我们使用模型个性化和标题修改来使我们的模型与目标领域保持一致，并发现与未保持一致的基线相比，我们的模型有所改进。我们使用模型个性化和标题修改来使我们的模型与目标领域保持一致，并发现与未对齐的基线相比有了改进。我们在 Pascal VOC 上训练的跨域物体检测模型在 Watercolor2K 上取得了 SOTA 的成绩。我们在城市景观上训练的跨域分割方法在黑暗苏黎世val和夜间驾驶上取得了SOTA结果。项目页面： vision.caltech.edu/TADP/ *代码页面：* github.com/damaggu/TADP



1. 导言

扩散模型在图像生成方面达到了最先进水平（SOTA）[32, 35, 38, 52]。最近，一些研究表明，经过扩散预训练的骨干模型在场景理解方面具有很强的先验性，这使得它们在高级判别视觉任务中表现出色，例如语义分割[17, 53]、单目深度估计[53]和关键点估计[28, 43]。我们将这些工作称为基于扩散的感知方法。与强制视觉语言模型（如 CLIP）不同 [22, 26, 31]、

*平等贡献。

图 1.**文本对齐扩散感知（TADP）。**在 TADP 中，图像标题将文本提示与传递给基于扩散的视觉模型的图像对齐。在跨领域任务中，目标主要信息被纳入提示中，以提高效率。

生成模型与文本有因果关系，文本引导图像生成。在潜像扩散模型中，文本提示会控制去噪 U-Net [36]，使图像潜像按照有语义的方向移动 [5]。

我们探索了这种关系，发现文本-图像对齐能显著提高基于扩散的感知性能。然后，我们研究了跨域视觉任务中的文本-目标域对齐，发现在源域训练的同时对齐目标域可以提高模型的目标域性能（图 1）。

我们首先研究了基于扩散的感知模型的提示，发现增加文本与图像的对齐度可以提高语义分割和深度估计的性能。

我们发现，未对齐的文本提示会给扩散模型的特征图带来语义偏移。我们发现，未对齐的文本提示会给扩散模型的特征图带来语义偏移。

[5]，而这些转移会增加特定任务头解决目标任务的难度。具体来说，我们要问的是，未对齐的文本提示（如平均特定类别的句子嵌入（[31, 53]））是否会通过交叉注意机制干扰特征图，从而阻碍表现。通过对 Pascal VOC2012 分割[14]和 ADE20K [55]的消融实验，我们发现偏离目标和缺失的类别名称会降低图像分割质量。我们的研究表明，自动图像标注[25]可实现文本与图像的充分对齐，从而达到感知效果。我们的方法（连同潜在表示缩放，见第 4.1 节）在 Pascal 和 ADE20k 上的语义分割性能分别提高了 4.0 mIoU 和 1.7 mIoU，在 NYUv2 [42] 上的深度估计性能提高了 0.2 RMSE（相对+8%），并设置了新的 SOTA。

接下来，我们将 重点放在跨域适应上：当模型在一个域中训练并在不同域中测试时，适当的图像字幕能否帮助视觉感知？使用适当的提示策略在源领域训练模型，可以在多个基准上实现出色的无监督跨领域性能。我们在 Pascal VOC [13, 14]上对我们的跨域方法进行了评估，结果显示，Watercolor2k (W2K) 和 Comic2k (C2K)

[21] 进行物体检测，从城市景观（CS）[9] 到苏黎世黑暗（DZ）[39] 和夜间（ND）驾驶[10] 进行语义分割。我们探索了不同程度的文本-目标域对齐方式，发现改进对齐方式会带来更好的性能。我们还使用两种扩散个性化方法进行了演示，即文本反转（Textual Inversion

[16]和 DreamBooth [37]，以获得更好的目标域对齐和性能。我们发现，在没有文本-目标域对齐的情况下，扩散预训练足以在所有跨域数据集上实现 SOTA（CS!DZ 上 +5.8 mIoU，CS!ND 上 +4.0 mIoU，VOC!W2k 上 +0.7 mIoU）或接近 SOTA 的结果，而加入我们的最佳文本-目标域对齐方法后，Wa-terolor2k 上的结果进一步提高了 +1.4 AP，Comic2k 上提高了 +2.1 AP，Night-time Driving 上提高了 +3.3 mIoU。

总体而言，我们的贡献如下

- 我们提出了一种自动生成标题的新方法，该方法通过提高文本与图像的对齐度，显著提高了多项基于扩散的视觉任务的性能。
- 我们系统地研究了提示如何影响基于扩散的视觉性能，阐明了类的存在、提示语中的语法以及之前使用的平均嵌入的影响。
- 我们证明，基于扩散的感知效果在不同领域中，文本-目标对齐可提高性能，而模型个性化可进一步提高性能。

2. 相关工作

2.1. 单域视觉任务的扩散模型

对扩散模型进行训练，以逆转逐步向前的噪声过程。一旦经过训练，它们就能从纯噪声中生成高度逼真的图像[32, 35, 38, 52]。为了控制图像生成，扩散模型是通过引导扩散过程的文本提示/说明来训练的。这些提示通过文本编码器生成文本嵌入，并通过交叉注意层纳入反向扩散过程。

最近，一些研究探索了在视觉分辨任务中使用扩散模型。具体做法可以是利用扩散模型作为任务的骨干[17, 28, 43, 53]，也可以是针对特定任务对扩散模型进行微调，然后利用它为下游模型生成合成数据[2, 50]。我们将扩散模型作为下游视觉任务的支柱。VPD [53] 将图像编码为潜在表征，并通过稳定扩散模型的一个步骤。交叉注意图、多尺度特征和输出潜码被串联起来，并传递到特定任务的头部。文本提示通过交叉注意机制影响所有这些映射，从而引导反向差异融合过程。交叉注意映射被纳入多尺度特征映射和输出潜在重现中。文本指导着扩散过程，并能相应地改变潜表征的语义矩阵[1, 5, 16, 18]。VPD 如何使用提示界面的细节将在第 3 节中介绍。简而言之，VPD 使用*未对齐的*文本提示。在我们的工作中，我们展示了如何通过使用字幕机将文本与图像对齐来显著改善语义分割和深度评估。信息性能。

2.2. 图片说明

CLIP[31]引入了一种新颖的学习范式，将图像与标题对齐。不久之后，包含 5B 图片-文本对的 LAION-5B 数据集[41]发布；该数据集被用于训练稳定扩散。我们推测，文本-图像对齐对于扩散预训练视觉模型非常重要。然而，在高级视觉任务（如分割和深度估计）中使用的图像并不能自然地与文字说明配对。为了获得图像对齐的标题，我们使用了 BLIP-2 [25]，这是一种反转 CLIP 潜在空间的模型，可为新图像生成标题。

2.3. 跨域视觉任务的扩散模型

有几项研究利用扩散模型对跨领域设置进行了探索[2, 17]。Benigmim 等人[2] 使用扩散模型为下游无监督适应（UDA）架构生成数据。在 [17] 中，扩散

主干被冻结，分割头在一致性损失的情况下进行训练，并通过类别和场景提示将潜码引向目标跨域。与 VPD 类似，类别提示包括数据集中所有类别的标记嵌入，而不管这些类别是否出现在任何特定图像中。一致性损失迫使模型为所有不同的场景提示预测相同的输出掩码，从而帮助分割头不受场景类型的影响。我们没有使用一致性损失，而是在源域数据上训练扩散模型骨干和任务头，同时在标题中加入或不加入目标域的风格。我们发现，更好地与目标域保持一致（即在提示中包含目标主要信息）会带来更好的跨域性能。

2.4. 跨域物体检测

根据训练/测试时可用的数据/标签，跨域物体检测可分为多个子类别。无监督域自适应异议检测（UDAOD）试图通过在无标签的目标域数据上进行训练来提高检测性能，采用的方法包括自我训练[11, 44]、对抗分布对齐[54]或为自我训练生成伪标签[23]。跨域弱监督对象检测（CDWSOD）假定在训练时有图像级注释，并利用伪标签[21, 30]、对齐[51]或对应关系挖掘[19]。最近，[46] 使用 CLIP [31] 进行单域泛化，旨在从单域泛化到多个未见的目标域。我们基于文本的方法是一种新的跨域对象检测方法，它试图从单一来源适应未见过的目标领域，只需将目标领域的广泛语义上下文（如雾/夜/漫画/水彩）作为我们方法的文本输入即可。当我们结合模型个性化时，我们的方法可被视为一种 UDAOD 方法，因为我们是基于未标记的目标图片来训练标记的。

3. 方法

稳定扩散 [35]。文本到图像的稳定扩散模型由四个网络组成：一个编码器 E、一个条件去噪自动编码器（稳定扩散中的 U-Net）
✓、语言编码器図✓（稳定扩散中的 CLIP 文本编码器）和解码器 D。
✓，使得 $D(E(x)) = x\tilde{}$ ↑ x。训练✏✓由预先定义的正向过程和学习的反向过程组成。反向过程是通过 LAION-400M [40]（一个包含 4 亿张图片（x 2 X）和标题（y 2 Y）的数据集）学习的。在正向处理过程中，图像 x 被编码成潜在 $z_0 = E(x)$，然后执行 t 步正向噪声处理过程，生成噪声潜在 z_t。然后将噪声潜在 z = E (x)，再执行 t 步向噪声处理过程，生成噪声潜在 z。

在反向过程中，潜在 z_t 与时间步长 t 和图像标题表示 $C = \boxtimes_\checkmark(y)$一起传递给去噪自动编码器 $\epsilon_\checkmark$。$\boxtimes_\checkmark$ 通过交叉关注机制将有关 y 的信息添加到 $\epsilon_\checkmark$，其中查询来自图像，键和值是标题表示的变换。训练模型$\epsilon_\checkmark$ 的目的是预测在前向过程的第 t 步中添加到潜变量中的噪声：

$$L_{LDM} := E_{E(x),y,\epsilon\leftarrow N(0,1),t}\left[\|\epsilon - \epsilon_\checkmark(z_t, t, \boxtimes_\checkmark(y))\|^2\right], \quad (1)$$

其中 $t \in \{0, ..., T\}$。在生成过程中，纯噪声 latent z_T 和用户指定的提示通过去噪自动编码器$\epsilon_\checkmark$ T 步，并解码 $D(z_0)$ 以生成由文本提示引导的图像。

用于特征提取的扩散。扩散骨架已在一些研究中用于下游视觉任务[17, 28, 43, 53]。由于其公开可用性和在感知任务中的表现，我们在 VPD 中使用了特征提取方法的修改版（见第 4.1 节）。图像潜在 $z_0 = E(x)$ 和条件 C 经过去噪过程的最后一步 $\epsilon_\checkmark(z_0, 0, C)$。U-Net 的交叉注意图 A 和多尺度特征图 F 被连接成 $V = A \oplus F$，并传递给特定任务的头部 H，以生成预测 $\hat{p} = H(V)$。骨干网$\epsilon_\checkmark$ 和头部 H 采用特定任务损失 $L_H(\hat{p}, p)$ 进行训练。

平均 EOS 标记。为了生成 C，以前的方法[17, 53]依赖于 CLIP [31] 的一种方法，即使用平均文本嵌入作为数据集中类别的表示。CLIP 文本编码器会为每个感兴趣的类别（如"<类名>的<形容词>照片"）提供一个包含 80 个句子模板的列表。我们用 B 来表示数据集中的类名集合。对于一个特定的类别（$b \in B$），CLIP 文本编码器会返回一个 $80 \times N \times D$ 张量，其中 N 是所有模板的最大标记数，D 是 768（每个标记嵌入的维度）。较短的句子用 EOS 标记填充，以填满最大标记数。对每个句子模板中的第一个 EOS 标记进行平均，并将其作为该类的代表性嵌入，这样 $C \in R^{|B|\times 768}$。这种方法在 [17, 53] 中使用，我们将其标记为 C_{avg}，并将其作为基线。对于语义分割，所有的类嵌入，无论是否出现在图像中，都会被传递到交叉注意层。只有房间类型的类嵌入才会传递给交叉注意层，用于深度估计。

3.1. 文本对齐扩散感知 (TADP)

我们的研究提出了一种新方法，用于提示扩散预训练感知模型。具体来说，我们探索了不同的提示方法 G 来生成 C。

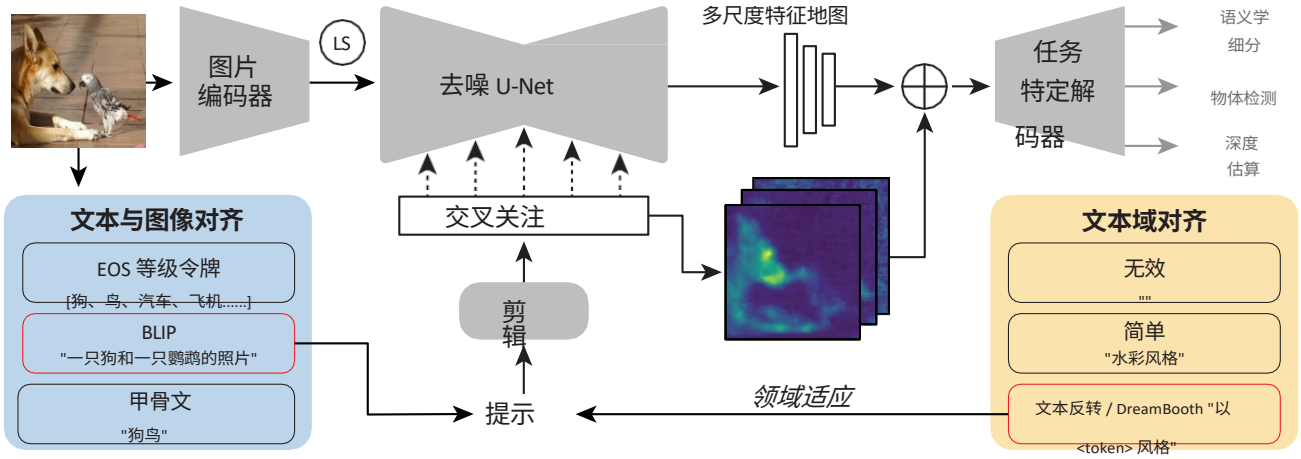


图 2.**TADP 概述。**我们测试了几种提示策略，并评估了它们对下游视觉任务性能的影响。我们的方法将交叉注意和多尺度特征图连接起来，然后再将它们传递给视觉专用解码器。在蓝色方框中，我们展示了三种具有不同文本图像对齐程度的单域字幕策略。我们建议使用 BLIP[25]字幕来提高图像-文本对齐度。我们将分析扩展到跨域设置（黄色方框），通过在源域生成的图像标题中添加标题修饰符，探索将源域文本标题与目标域对齐是否会影响模型性能，并发现模型个性化修饰符（文本反转/梦境）效果最佳。

该模型使用 BLIP-2 [25]（一种图像标题算法）生成标题，作为模型的条件：$G(x) = y$ ~ *！*然后，我们通过将目标域信息纳入 $C = C + M(P)_s$，将我们的方法扩展到跨域设置，其中 M 是字幕修改器，它将目标域信息 P 作为输入，并输出字幕修改器 $M(P)_s$ 和模型修改器 $M(P)_{\checkmark}$。在第 4 节中，我们将通过系统地改变字幕器 G 和字幕修改器 M *来*分析扩散模型的文本-图像接口，*以完成*三种不同的视觉任务：语义分割、物体检测、

和单目深度估计。我们的方法和实验见图 2。按照文献[53]，我们使用快速和常规计划训练 ADE20k 分割和 NYUv2 深度估计模式。在 ADE20k 上，我们使用 4k 步（快速）、8k 步（快速）和 80k 步（正常）进行训练。对于 NYUv2 深度，我们使用 1 个波段（快速）和 25 个波段（正常）进行训练。有关实施细节，请参阅附录 D。

方法	平均值	电讯	LS	G	加时赛	mIoU^{ss}
VPD(R) [53]	X	X			X	82.34
VPD(LS)	X	X	X		X	83.06
班级标志			X		X	82.72
班级名称			X		X	84.08
TADP-0			X	X		86.36
TADP-20			X	X		86.19
TADP-40			X	X		**87.11**
TADP(NO)-20			X			86.35
甲骨文 TADP			X			85.85

我们发现，使用 C_{avg} 进行分割和深度估计时，潜在缩放提高了性能（图 3）。具体来说，潜在缩放在 Pascal 上提高了 $\leftarrow 0$.8% mIoU，在 ADE20K 上提高了 $\leftarrow 0$.3% mIoU，在 NYUv2 Depth 上提高了相对 $\leftarrow 5$.5% RMSE（图 3）。

4. 成果

4.1. 潜在缩放

在探索图像-文本配准之前，我们先对编码图像进行潜在缩放（Rombach 等人[35]的附录 G）。这将图像潜影归一化为标准正态分布。缩放因子固定为 0.18215。

表 1.**Pascal VOC2012 分段的提示。**我们报告了 Pascal 实验的单尺度验证 mIoU。(R）：再现 VPD，Avg：EOS 标记平均，LS：La- tent Scaling，G：Grammar，OT：Off-target information。对于我们的方法，我们用 TADP-X 表示 BLIP 标题的最小长度，用 (NO) 表示名词的最小长度。

4.2. 单域对齐

平均 EOS 令牌。我们仔细研究了 C 的平均 EOS 标记的使用（见第 3 章）。在测量 CLIP 潜在空间中的余弦相似性时，平均 EOS 标记是合理的，但在扩散模型中它并不适用，因为在扩散模型中，文本通过交叉注意引导扩散过程。在定性分析中，我们发现平均 EOS 标记会降低交叉注意图的质量（图 4）。相反，我们探索使用 CLIP 将每个类名独立嵌入，并使用与实际单词（而非 EOS 标记）相对应的标记，将其作为交叉注意层的输入：

$$_{\text{GClassEmbs}}(B) = \text{concat}(CLIP(\text{b})|\text{b } 2\ B)\ !\ _{\text{CClassEmbs}}\ (2)$$

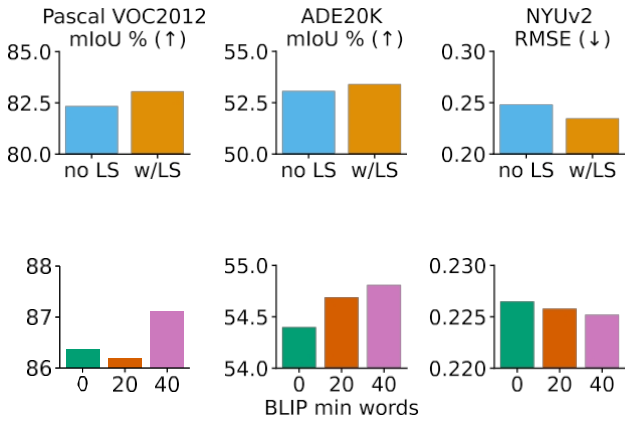


图 3.**Latent Scaling (LS) 和 BLIP caption 最小长度的影响。**我们报告了 Pascal 的 mIoU、ADE20K 的 mIoU 和 NYUv2 深度的 RMSE（右图）。(上图）潜在缩放提高了 Pascal ↼0.8 mIoU（越高越好）、↼0.3 mIoU 和↼5.5% 相对 RMSE（越低越好）的性能。(下图）我们看到 BLIP 最小标记长度也有类似的效果，较长的字幕表现更好，在 Pas- cal 上提高了 ↼0.8 mIoU，在 ADE20K 上提高了 ↼0.9 mIoU，相对 RMSE 提高了 ↼0.6%。

其次，我们探索通用提示符，即一串用空格分隔的类名：

$$G_{\text{ClassNames}}(B) = \{\text{''}+ \text{b|b } 2\ B\} \ ! \ C_{\text{ClassNames}} \qquad (3)$$

这些提示与用于平均 EOS 标记 C_{avg} 的提示类似，但使用的是与代表类别名称的单词相对应的标记。我们在 Pascal VOC2012 分段中对这些变体进行了评估。我们发现，$_{\text{CClassNames}}$ 将性能提高了 1.0 mIoU，而 $_{\text{CClassEmbs}}$ 则将性能降低了 0.3 mIoU（见表 1）。我们在附录 A 中更深入地分析了文本-图像对齐对不同融合模型的交叉注意图和图像生成特性的影响。

TADP。为了将扩散模型文本输入与图像对齐，我们使用 BLIP-2 [25] 为单域数据集（Pascal、ADE20K 和 NYUv2）中的每个图像生成标题。

$$G_{\text{TADP}}(\text{x}) = \textit{BLIP-2}(\text{x}) \ ! \ C_{\text{TADP}}(\text{x}) \qquad (4)$$

BLIP-2 的训练目的是生成图像对齐的文字说明，它是围绕 CLIP 潜在空间设计的。不过，也可以使用其他能生成标题的视觉语言算法。我们发现，这些文字说明提高了所有数据集和任务的性能（表 1、2、3）。帕斯卡分割的性能提高了 ↼4% mIoU，ADE20K 的性能提高了 ↼1.4% mIoU，NYUv2 深度的相对 RMSE 提高了 4%。我们发现，ADE20K 对快速时间表的影响更大，提高了

5 mIoU（4K），↼2.4 mIoU（8K）。在 NYUv2 深度上，我们看到快速计划的增益较小↼2.4%。报告的所有数字都是相对于 VPD 的潜在缩放。

方法	#Params	FLOPs	Crop	mIoU^{ss}	mIoU^{ms}
自我监督预训练					

VPD(LS)	862M	891G	512^2	53.7	54.4
TADP-40 （我们的）	862M	2168G	512^2	**54.8**	**55.9**
甲骨文TADP	862M	-	512^2	72.0	-

表 2. **使用不同方法对 ADE20k 进行语义分割。**我们的方法（绿色）在扩散预训练模型类别中获得了 SOTA。我们的方法（绿色）在扩散预处理模型类别中达到了 SOTA，这表明基于扩散的模型具有未来研究的潜力，因为它明显高于总体 SOTA（黄色）。参见表 1 中的符号说明和表 2 中的符号说明。符号说明见表 1，快速时间表结果见表 S1。S1 为快速计划结果。

方法	*RMSE#*	δ_1 "	δ_2 "	δ_3 "	REL #	log10 #
默认时间表						
SwinV2-L [27]	0.287	0.949	0.994	0.999	0.083	0.035
AiT [29]	0.275	0.954	0.994	0.999	0.076	0.033
ZoeDepth [3]	0.270	0.955	0.995	0.999	0.075	0.032
VPD [53]	0.254	0.964	0.995	0.999	0.069	0.030
VPD(R)	0.248	0.965	0.995	0.999	0.068	0.029
VPD(LS)	0.235	0.971	0.996	0.999	0.064	0.028
TADP-40	**0.225**	**0.976**	**0.997**	**0.999**	**0.062**	**0.027**
快速时间表，1 个纪元						
VPD	0.349	0.909	0.989	0.998	0.098	0.043
VPD(R)	0.340	0.910	0.987	0.997	0.100	0.042
VPD(LS)	0.332	0.926	0.992	0.998	0.097	0.041
TADP-0	**0.328**	**0.935**	**0.993**	**0.999**	**0.082**	**0.038**

表 3.**NYUv2 中的深度估计。**我们发现潜在缩放在 RMSE 指标上的相对增益为 ↼ 5.5%。通常情况下，图像-文本对齐在 RMSE 指标上相对提高↼ 4%。我们还探索在 TADP 中添加文本适配器 (TA)，但没有发现明显的增益。符号说明见表 1。

我们进行了一些删减，以分析字幕的哪些方面比较重要。我们探讨了 BLIP-2 的最小标记数超参数，以探索较长的字幕是否能为顺流任务生成更有用的特征图。我们尝试了 0、20 和 40 个标记的最小标记数（记为 $_{\text{CTADP-N}}$），发现较长字幕的增益较小但一致，40 个标记与 0 个标记相比，平均相对增益为 0.75%（图 3）。接下来，我们对 Pascal $_{\text{CTADP-20}}$ 标题进行了消减，以了解 Pascal $_{\text{CTADP-20 标题}}$中的

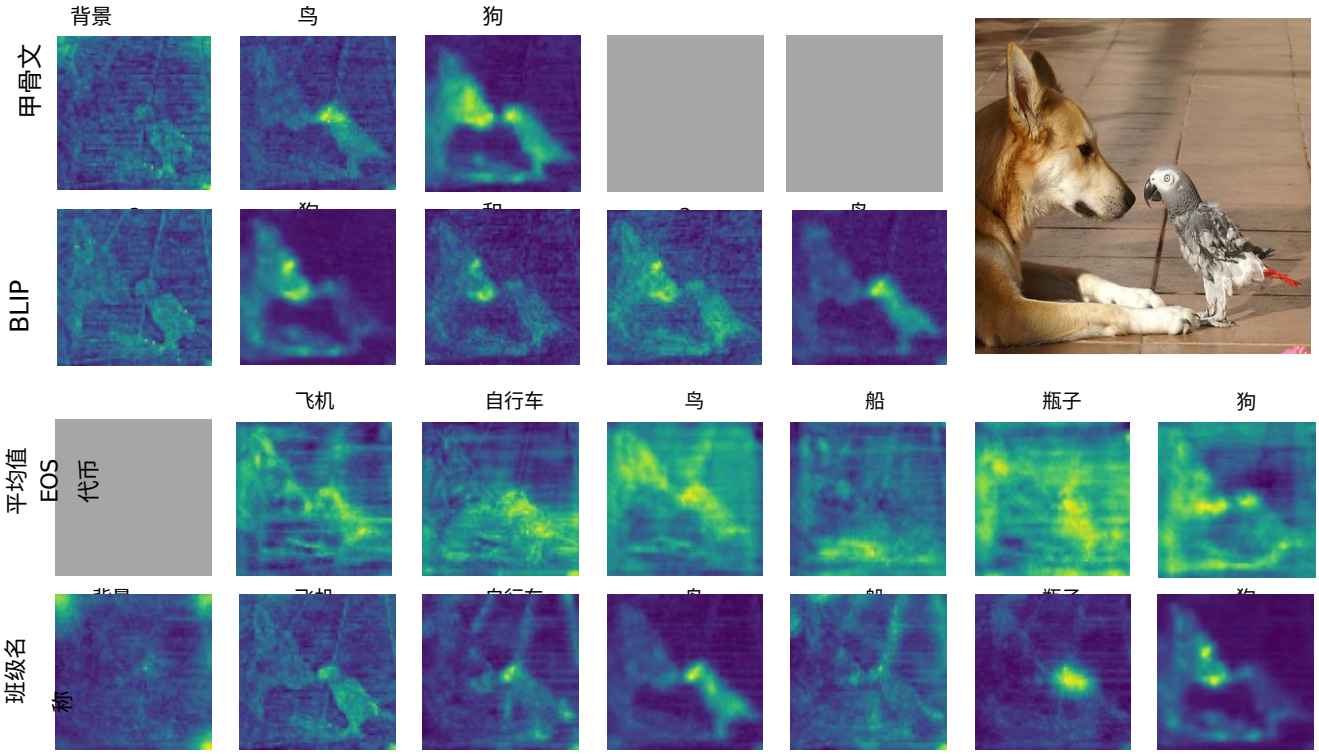


图 4.**不同类型提示的交叉注意图（训练前）。**我们比较了四种提示类型的交叉注意图：oracle、BLIP、平均 EOS 标记和以空格分隔的字符串表示的类名。所有不同尺度的不同头部的交叉注意图都被上采样为 64x64 并取平均值。在比较平均模板 EOS 和类名时，我们发现（从质量上）平均化会降低交叉注意图的质量。此外，我们还发现，图像中不存在的类别名称也可能具有高度局部化的注意图（例如 "瓶子"）。关于交叉注意图的进一步分析，我们将在 A 节中探讨图像到图像生成、复制粘贴图像修改等问题。

要实现我们观察到的性能提升，就必须对字幕中的名词进行过滤。我们使用 NLTK [4] 过滤字幕中的名词。在 $_{\text{CTADP(NO)-20}}$ 仅有名词的字幕设置中，我们实现了 86.4% 的 mIoU，与 $_{\text{CTADP-20}}$ 的 86.2% mIoU 相似（表 1），这表明有名词就足够了。

甲骨文关于名词的这一见解让我们产生了这样的疑问：甲骨文标题，即图像中的所有对象类别名称都作为标题提供，是否能进一步提高性能？我们将 $B(x)$ 定义为图像 x 中出现的类别名称集合。

$$_{\text{GOracle}}(x) = \{''+ b|b\ 2\ B(x)\}\ \ !\ _{\text{COracle}}(x) \qquad (5)$$

虽然这不是一个现实的设置，但它可以作为我们的方法在分割任务上的性能的近似上限。我们发现，该方法在分割方面的性能有很大提高，在 Pascal 和 ADE20K 上分别达到了 89% 和 72.2% 的 mIoU。在深度估计方面，我们只为较小的图像子集提供了多类分割掩码，因此无法生成可比较的算法。我们对神谕标题进行消减，以评估模型对排列的敏感性。对于 ADE20K，在 4k 迭代计划中，我们修改了神谕标题

通过随机添加和删除类别，使再召回率和精确率分别为 0.5、0.75 和 1.0（各不相同）（表 S2）。我们发现精确度和召回率都有影响，但召回率明显更重要。当召回率较低时（0.50），提高精确度的影响最小（<1% mIoU）。然而，当召回率增加到 0.75 和 1.00 时，精确度的影响逐渐增大 *(↞3%* mIoU 和 ↞7% mIoU）。相比之下，召回率在每个精确度级别都有很大影响：0.5 - *(↞6%* mIoU)、0.75 - *(↞9% mIoU)* 和 1.00 - *(↞13%* mIoU)。BLIP-2 的字幕精确度为 1.00，召回率为 0.5（表 2）。有关精确度、召回率和对象大小的其他分析见附录 B。

4.3. 跨域对齐

接下来，我们要问的是，文本-图像配准是否有利于跨领域任务。在跨领域任务中，我们在源领域训练一个模型，然后在不同的目标领域进行测试。跨域对齐有两个方面：第一个方面在单域中也存在，即图像-文本对齐；第二个方面是跨域中独有的，即文本-目标域对齐。第二个方面

方法	黑暗苏黎世-val mIoU	ND mIoU
DAFormer [20]	-	54.1
Refign-DAFormer [7]	-	56.8
PTDiffSeg [17]	37.0	-
$TADP_{null}$	**42.8**	57.5
TADPsimple	39.1	56.9
TADPTextualInversion	41.4	**60.8**
TADPDreamBooth	38.9	60.4
TADPNearbyDomain	41.9	56.9
TADPUnrelatedDomain	42.3	55.1

表 4.**跨域语义分割。**从城市景观（CD）到黑暗苏黎世（DZ）val 和夜间驾驶（ND）。我们报告的是 mIoU。我们的方法为 DarkZurich 和 Nighttime Driving 设置了新的 SOTA。

方法	水彩 2k		Comic2k	
	AP	AP50	AP	AP50
单域泛化（SGD）				
夹住缝隙 [46]	-	33.5	-	43.4
跨域弱监督物体检测				
PLGE [30]	-	56.5	-	41.7
ICCM [19]	-	57.4	-	37.1
H2FA R-CNN [51]	-	59.9	-	46.4
无监督领域适应性目标检测				
ADDA [45]	-	49.8	-	23.8
MCAR [54]	-	56.0	-	33.5
UMT [11]	-	58.1	-	-
DASS-Detector （额外数据）	-	71.5	-	**64.2**
TADPnull	42.1	72.1	31.1	**57.4**
TADPsimple	**43.5**	**72.2**	31.9	56.6
TADPTextualInversion	43.2	**72.2**	**33.2**	**57.4**
TADPDreamBooth	43.2	**72.2**	32.9	56.9
TADPNearbyDomain	42.0	71.5	31.8	56.4
TADPUnrelatedDomain	42.2	71.9	32.0	55.9

标题进行修改。

由于源域和目标域之间存在着很大的域偏移，因此这具有挑战性。我们的直觉是，虽然模型无法从训练图像中获得目标域的信息，但适当的文本提示可能会携带一些关于目标域的一般信息。我们的跨域实验侧重于文本-目标域对齐，并使用 G_{TADP} 进行图像-文本对齐（遵循我们从单域设置中获得的启示）。

训练我们在这种情况下的实验设计如下：我们在源域字幕 $C_{\text{TADP}}(x)$ 上训练扩散模型。通过这些源域标题，我们实验了四种不同的标题修改（每种修改都会增加对焦域的对齐度），一种是空的 $M_{null}(P)$ 标题修改，其中 $M_{null}(P)_s = ø = M_{null}(P)_{✏\surd} = ø$、一个简单的 $_{Msimple}(P)$ 标题修改器，其中 $_{Msimple}(P)_s$ 是一个手工制作的字符串，描述目标域的风格，并附加在末尾，$_{Msimple}(P)_{✏\surd} = ø$，一个文本反转[16] $M_{TI}(P)$ 标题修改器，其中输出 $M_{TI}(P)_s$ 是一个学习的文本反转标记 <*> 和 $M(P) = ø$。<*> 和 $M_{TI}(P)_{✏\surd} = ø$，以及 DreamBooth [37] $M_{DB}(P)$ 字幕修改器，其中 $M_{DB}(P)_s$ 是学习的 DreamBooth 标记 <SKS>，$M_{DB}(P)_{✏\surd}$ 是 DreamBoothed 扩散骨干。我们还包括两个额外的对照实验。在第一个实验中，$M_{ud}(P)$ 在字符串末尾添加了一个*不相关的*目标域样式。在第二个实验中，$M_{nd}(P)$在标题后附加了一个*邻近*但不同的目标域样式。与其他方法相比，$M_{TI}(P)$ 和 $M_{DB}(P)$ 需要更多的信息，例如 P 代表目标域中未标记图像的子集。

测试。在焦油域图像上测试训练有素的模型时，我们希望测试图像的标题修改与训练设置相同。但是，G_{TADP} 会带来一个干扰，因为它自然地会对测试图像的

表 5.**跨域对象检测。从** Pascal VOC 到 Water- color2k 和 Comic2k。我们报告了 AP 和 $_{AP50}$。我们的方法为 Watercolor2K 设定了新的 SOTA。

反弹包含目标域信息。例如，对于来自 Water-color2K 数据集的图像，G_{TADP} (x) 可能会生成 "一只狗和一只鸟的水彩画 "的标题。对这一提示使用 $_{Msimple}(P)$ 标题修改器会引入冗余信息，并且与训练时使用的标题格式不匹配。为了移除目标域信息，并得到一个可按与训练数据相同的方式修改的纯标题，我们使用 GPT-3.5 [6] 移除所有提及目标域的移位。例如，在使用 GPT-3.5 删除上述句子中提到的水彩风格后，我们只剩下 "一只鸟和一只狗的图像"。有了这些*经过 GPT-3.5 净化的标题*，我们就可以在评估测试图像时匹配训练时使用的标题修改。通过这种标题清理策略，我们可以控制测试图像标题中目标域信息的包含方式，确保测试标题与训练标题处于同一域。

评估。我们在多个数据集上对跨域转移进行了评估。我们在 Pascal VOC [13, 14] 对象检测上训练模型，并在 Watercolor2K (W2K) [21] 和 Comic2K (C2K) [21] 上进行评估。我们还在 Cityscapes [9] 数据集上训练模型，并在 Nighttime Driv- ing (ND) [10] 和 Dark Zurich-val (DZ-val) [39] 数据集上进行评估。结果见表 4、5。4, 5.在接下来的章节中，我们还将报告每种方法在跨域分割数据集（平均 mIoU）和跨域对象检测数据集（平均 AP）上的平均性能。

空标题修饰符。空字幕没有目标域信息。在这种情况下，模型使用没有目标域信息的字幕进行训练，并使用 GPT-3.5 净化过的目标域字幕进行测试。我们发现，仅使用普通字幕（无目标域信息）进行扩散预训练本身就异常强大；该模型已在 VOC!W2K 上实现了 SOTA（72.1 AP_{50} ），在 CD!DZ-val 上实现了 SOTA（42.8 mIoU），在 CD!ND 上实现了 SOTA（60.8 mIoU）。我们的模型在 VOC！W2K 上的性能优于当前的 SOTA [44]，而在 VOC！C2K 上则较差（表 5 中用黄色标出）。不过，[44] 使用了来自目标（漫画）领域的大量额外训练数据集，因此我们在表 5 中用粗体标出了我们的结果，以显示它们优于所有 SOTA 模型。因此，我们在表 5 中用粗体标出了我们的结果，以显示它们优于所有其他只使用 C2K 中的图像作为目标领域示例的方法。此外，这些结果是在轻量级 FPN [24]头，而其他竞争方法，如 Re- fign [7]，则使用更重的解码器头。这些字幕的平均 mIoU 值为 50.5，平均 AP 值为 36.6。

简单字幕修改器。然后，我们在通用标题前加上目标域的语义转换，从而为我们的标题添加目标域的内容。这些字幕修改器都是手工制作的。例如，"一只狗和一只鸟 "就是 "一幅 X 风格的狗和鸟的画"（在 W2K 中，X 是水彩画，在 C2K 中，X 是漫画），而在 DZ 中，"一只狗和一只鸟的暗夜照片 "就是 "一幅 X 风格的狗和鸟的画"。这些标题实现了
48.0 个平均 mIoU 和 37.7 个平均 AP。

文字反转字幕修改器。文本反转 [16] 是一种从一组图像中学习目标概念（对象或风格）并将其编码为新标记的方法。我们从目标域图像样本中学习新的标记，以进一步提高图像与文本的对齐度（详见 D.1 节）。在这种情况下，句子模板是 "一幅<标记>风格的狗和鸟的画"。我们发现，平均而言，文本反转标题的表现最好，平均 mIoU 为 51.1，平均 AP 为 38.2。

梦ooth 字幕修改器。DreamBooth-ing [37] 旨在实现与文本反转相同的目标。在学习新标记的同时，我们会使用一组目标域图像对稳定扩散骨干网进行微调（细节见 D.1 节）。在训练之前，我们将稳定扩散骨干网与 DreamBooth 编辑的骨干网进行交换。我们使用与文本反转相同的模板。这些字幕的平均 mIoU 为 49.7，平均 AP 为 38.1。

消减。我们通过引入无关和邻近的目标域风格修改来消减我们的目标域对齐策略。例如，对于 W2K 和 C2K 数据集，这将是 "一张狗和一只鸟的**数码相机**照片"（不相关）和 "一幅狗和一只鸟的**建筑**画"（附近）。在 ND 和 DZ-val 数据集中，"街道上一辆汽车的**水彩**画"（不相关）和 "街道上一辆汽车的**雾景**照片"。我们发现这些非目标域降低了所有数据集的性能。

5. 讨论

我们提出了一种通用、全自动的图像-文本配准方法，可应用于任何基于扩散的感知模型。为此，我们系统地探讨了文本-图像对齐对语义分割、深度估计和物体检测的影响。我们研究了类似的原理是否适用于跨领域设置，并发现在训练过程中向目标领域对齐可提高下游跨领域性能。

我们发现，用于提示的 EOS 标记平均法不如用于图像对象的字符串法有效。我们的甲骨文消融实验表明，我们的困难预训练分割模型对缺失类别特别敏感（召回率降低），而对偏离目标类别不太敏感（精度降低），两者都有负面影响。我们的结果表明，将文本提示与图像对齐对于为下游分割头识别/生成良好的多尺度特征图非常重要。这意味着，在扩散模型中，如果没有文本的指导，多尺度特征和潜在代表无法自然识别语义概念。此外，适当的潜在缩放对于下游视觉任务至关重要。最后，我们展示了如何使用具有开放词汇、高精度和与下游任务无关等优点的字幕机来提示扩散预训练分割模型，从而自动提高性能，而不是提供所有可能的类名。

我们还发现，扩散模型可以有效地用于跨领域任务。由于扩散骨干的泛化能力，我们的模型在没有任何字幕的情况下，已经在跨领域任务中超越了几项 SOTA 结果。我们发现，良好的目标域对齐有助于提高某些域的跨域性能，而错误对齐则会导致性能下降。仅从词语中捕捉目标域风格的形成可能比较困难。对于这些情况，我们表明，通过文本反转或 Dreambooth 进行模型标注可以弥补这一差距，而无需标注数据。未来的工作可以探索如何将我们的框架扩展到多个未见领域。未来的工作还可以探索更针对特定任务的封闭词汇字幕机，以接近甲骨文级别的性能。

致谢。皮埃特罗-佩罗纳和马库斯-马克斯得到了美国国立卫生研究院（NIH R01 MH123612A）和加州理工学院陈研究所（Neuroscience Re- search Grant Award）的资助。Pietro Perona、Neehar Kondapaneni、Roge- rio Guimaraes和Markus Marks得到了西蒙斯基金会（NC-GB-CULM-00002953-02）的资助。曼努埃尔-诺特得到了苏黎世联邦理工学院博士流动奖学金的资助。我们感谢 Oisin Mac Aodha、Yisong Yue 和 Mathieu Salzmann 的宝贵意见，这些意见有助于改进本研究工作。

参考资料

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Ji- aming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu.2

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalo- geiton, and Ste´phane Lathuilie`re.使用个性化扩散模型的单次无监督领域适应。*arXiv 预印本 arXiv:2303.18080*，2023。2

[3] Shariq Farooq Bhat、Reiner Birkl、Diana Wofk、Peter Wonka 和 Matthias Mu¨ller.ZoeDepth： *ArXiv preprint arXiv:2302.12288*, 2023.5

[4] Steven Bird、Ewan Klein 和 Edward Loper。*用 Python 处理自然语言：用自然语言工具包分析文本》*。O'Reilly Media 公司，2009 年。6

[5] Manuel Brack、Felix Friedrich、Dominik Hintersdorf、Lukas Struppek 、 Patrick Schramowski 和 Kristian Kersting。SEGA：使用语义维度指导扩散。*arXiv 预印本 arXiv:2301.12247*, 2023.1, 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners.*神经形成处理系统进展》*，33:1877-1901，2020 年。7

[7] David Bru¨ggemann、Christos Sakaridis、Prune Truong 和 Luc Van Gool。Refign：根据不利条件调整语义分割的对齐和提炼。*2023 年IEEE/CVF 计算机视觉应用冬季会议（WACV）*，2022年。7, 8

[8] 陈哲、段雨辰、王文海、何俊军、卢彤、戴继峰和 Y. Qiao.乔。用于密集预测的视觉变换器适配器。*arXiv 预印本 arXiv:2205.08534*, 2022.5

[9] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、Stefan Roth 和 Bernt Schiele。用于城市场景语义理解的城市景观数据集》。*2016 IEEE 计算机视觉与模式识别大会（CVPR）*，第 3213-3223 页，2016 年。2, 7

[10] Dengxin Dai 和 Luc Van Gool.黑暗模型适应：从白天到夜晚的语义图像分割。*2018 第 21 届智能交通系统国际会议（ITSC）*，第 3819-3824 页，2018 年。2, 7

[11] 邓金红、李文、陈玉华、段立新。用于跨域物体检测的无偏均值教师。*IEEE/CVF 计算机视觉与模式识别会议论文集》*，第 4091-4101 页，2021 年。3, 7

[12] David Eigen、Christian Puhrsch 和 Rob Fergus。使用多尺度深度网络从单张图像预测深度图。*神经信息处理系统进展 27 (NIPS 2014)*，2014 年。14

[13] M.Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman.PASCAL 可视化对象类

2007 年挑战赛（VOC2007）结果。http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html。2, 7

[14] M.Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A.Zisserman.PASCAL Visual Object Classes Challenge 2012 (VOC2012), 2012.2, 7

[15] 方玉新、王文、谢斌辉、孙全森、吴乐玉、王兴刚、黄铁军、王新龙和曹悦。EVA：探索规模化遮蔽视觉表征学习的极限。*2023 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 19358-19369 页，2022 年。5

[16] Rinon Gal 、 Yuval Alaluf 、 Yuval Atzmon 、 Or Patashnik、Amit H. Bermano、Gal Chechik 和 Daniel Cohen-Or。一图一字：使用文本反转实现文本到图像的个性化生成。*arXiv 预印本 arXiv:2208.01618*, 2022.2, 7, 8

[17] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Man- gas, and Luc Van Gool.用于跨域语义分割的提示扩散描述（Prompting Diffusion Represen- tations for Cross-Domain Semantic Segmentation ） 。 *arXiv preprint arXiv:2307.02138*, 2023.1, 2, 3, 7

[18] 阿米尔-赫兹、罗恩-莫卡迪、杰伊-特南鲍姆、克菲尔-阿伯曼、雅伊尔-普里奇和丹尼尔-科恩-奥尔。带交叉注意控制的提示到提示 Im- age 编辑。*arXiv 预印本 arXiv:2208.01626*, 2022.2

[19] Luwei Hou, Yu Zhang, Kui Fu, and Jia Li.用于跨域弱监督对象检测的信息一致对应挖掘。*IEEE/CVF 计算机视觉与模式识别会议论文集》*，第 9929-9938 页，2021 年。3, 7

[20] Lukas Hoyer 、 Dengxin Dai 和 Luc Van Gool 。DAFormer：改进领域自适应语义分割的网络架构和训练策略。*2022 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 9914-9925 页，2022 年。7

[21] 井上直人（Naoto Inoue）、古田良介（Ryosuke Furuta）、山崎俊彦（Toshihiko Yamasaki）和相泽清治（Kiyoharu Aizawa）。通过渐进式域自适应进行跨域弱监督物体检测。In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001-5009, 2018.2, 3, 7

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig.利用噪声文本监督扩展视觉和视觉语言表述学习》，*arXiv preprint arXiv:2102.05918*, 2021.1

[23] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long.跨域对象检测的解耦适应。*arXiv 预印本 arXiv:2110.02578*, 2021.3

[24] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dolla´r.全景特征金字塔网络。*2019 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 6392-6401 页，2019 年。8, 14

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi.BLIP-2：使用冻结图像编码器和大型语言模型进行语言图像预训练的引导（Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models）。*ArXiv 预印本 arXiv:2301.12597*, 2023.2, 4, 5

[26] 李阳光、梁峰、赵立晨、崔玉峰、欧阳万里、邵晶、于凤伟、闫俊杰。随处可见：*ArXiv preprint arXiv:2110.05208*, 2022.1

[27] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo.Swin Transformer V2：提升容量和分辨率。*2022 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 11999-12009 页，2021 年。5

[28] Grace Luo、Lisa Dunlap、Dong Huk Park、Aleksander Holyn- ski 和 Trevor Darrell。扩散超特征：*ArXiv preprint arXiv:2305.14334*, 2023.1, 2, 3

[29] Jia Ning, Chen Li, Zheng Zhang, Zigang Geng, Qi Dai, Kun He, and Han Hu. 全靠代币：*ArXiv preprint arXiv:2301.02229*, 2023.5

[30] 欧阳胜雄、王兴禄、柳克杰、李英明。用于跨域弱监督物体检测的伪标签生成-评估框架。In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 724-728.IEEE, 2021.3, 7

[31] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可转移的视觉模型。*国际机器学习大会*，第 8748-8763 页，2021 年。1, 2, 3

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.使用 CLIP Latents 进行分层文本条件图像生成。*arXiv 预印本 arXiv:2204.06125*, 2022.1, 2

[33] 饶永明、赵文亮、陈广义、唐岩松、朱铮、黄冠、周杰和卢继文。DenseCLIP：具有语境感知提示的语言引导密集预测。*2022 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 18061-18070 页，2022 年。5

[34] 任少卿、何开明、罗斯-吉尔希克和孙健。更快的 R-CNN：利用区域建议网络实现实时物体检测。*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137-1149, 2017.
14

[35] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bjorn Ommer。利用潜在扩散模型合成高分辨率图像。*2022 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 10674-10685 页，2022 年。1, 2, 3, 4

[36] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox：用于生物医学图像分割的卷积网络。*医学影像计算与计算机辅助干预 - MICCAI 2015》*，第 234-241 页。
Springer International Publishing, Cham, 2015.1

[37] 纳塔尼尔-鲁伊斯、李元珍、瓦伦-詹帕尼、雅伊尔-普里奇、迈克尔-鲁宾斯坦和克菲尔-阿伯曼。梦幻展台：精美

为主题驱动生成调整文本到图像扩散模型》，*arXiv preprint arXiv:2208.12242*, 2022.2, 7, 8

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi.具有深度语言理解能力的逼真文本到图像扩散模型。*arXiv 预印本 arXiv:2205.11487*, 2022.1, 2

[39] Christos Sakaridis、Dengxin Dai 和 Luc Van Gool。用于语义夜间图像分割的引导式课程模型适应和不确定性感知评估。*2019 IEEE/CVF 国际计算机视觉大会（ICCV）*，第 7373-7382 页，2019 年。2, 7

[40] Christoph Schuhmann 、 Richard Vencu 、 Romain Beaumont、 Robert Kaczmarczyk、 Clayton Mullis、 Aarush Katta、 Theo Coombes、 Jenia Jitsev 和 Aran Komatsuzaki。LAION- 400M：经 CLIP 滤波处理的 4 亿图像的开放数据集。
ArXiv preprint arXiv:2111.02114, 2021.3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Worts- man, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Je- nia Jitsev.LAION-5B：用于训练下一代图像-文本模型的开放式大规模数据集。*arXiv 预印本 arXiv:2210.08402*, 2022.2

[42] Nathan Silberman、Derek Hoiem、Pushmeet Kohli 和 Rob Fergus。从 RGBD 图像进行室内分割和支持推理。*欧洲计算机视觉会议（ECCV）*，2012 年。2

[43] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan.来自图像扩散的新兴对应关系》，*arXiv 预印本 arXiv:2306.03881*, 2023.1, 2, 3

[44] Barıs, Batuhan Topal, Deniz Yuret, and Tevfik Metin Sez- gin.图纸中人脸和人体检测的 领域自适应自监督预训练》。*ArXiv 预印本 arXiv:2211.10641*, 2022.3, 7, 8

[45] Eric Tzeng、Judy Hoffman、Kate Saenko 和 Trevor Darrell。对抗性判别域适应。In *Proceed- ings of the IEEE conference on computer vision and pattern recognition*, pages 7167-7176, 2017.7

[46] Vidit Vidit、Martin Engilberge 和 Mathieu Salzmann。CLIP the Gap: A Single Domain Generalization Approach for Ob- ject Detection.*2023 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 3219-3229 页，2023 年。3, 7

[47] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou.ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *ArXiv preprint arXiv:2305.11172*, 2023.5

[48] 王文海、戴继峰、陈哲、黄振航、李志奇、朱锡洲、胡晓华、卢彤、卢乐伟、李红生、王晓刚和乔颖。Qiao.InternIm- age：探索大规模视觉基础模型

可变形卷积。*2023 年 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 14408-14419 页，2022 年。5

[49] 王文、鲍杭波、董力、约翰-比约克、彭志亮 、 刘强波 、克里蒂-阿加瓦尔、欧瓦伊斯-汗-穆-哈迈德、萨克沙姆-辛格哈尔、苏博吉特-索姆、魏福如。图像作为外语：视觉和视觉语言任务的 BEIT 预训练。*2023 IEEE/CVF 计算机视觉与模式识别大会（CVPR）*，第 19175-19186 页，2023 年。5

[50] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen.DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models. *ArXiv preprint arXiv:2303.11681*, 2023.2

[51] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang.H2fa r-cnn：用于跨域弱监督物体检测的整体和分层特征配准。*IEEE/CVF 计算机视觉与模式识别会议论文集》*，第 14329-14339 页，2022 年。3, 7

[52] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchin- son, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu.用于内容丰富的文本到图像生成的比例自回归模型。1, 2

[53] 赵文亮、饶永明、刘祖彦、刘本林、周杰、卢继文。释放视觉感知中的文本-图像差异融合模型》，*arXiv preprint arXiv:2303.02153*, 2023.1, 2, 3, 4, 5, 14

[54] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye.双多标签预测的自适应物体检测。In *Computer Vision-ECCV 2020: 16th European Confer- ence, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII 16*, pages 54-69.Springer, 2020.3, 7

[55] 周博磊、赵航、泽维尔-普伊格、萨尼亚-菲德勒、阿德拉-巴里乌索和安东尼奥-托拉尔巴。通过 ADE20K 数据集进行场景解析。*2017 IEEE 计算机视觉与模式识别大会（CVPR）*，第 5122-5130 页，2017 年。2