(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117392675 A
(43)申请公布日 2024.01.12

(21)申请号 202311293866.1

(22)申请日 2023.10.09

(71)申请人 杭州电子科技大学
地址 310018 浙江省杭州市钱塘区白杨街道2号大街1158号

(72)发明人 陈鼎 王可逸 余宙 俞俊

(51)Int.Cl.
G06V 20/70(2022.01)
G06V 40/20(2022.01)
G06V 10/774(2022.01)
G06V 10/82(2022.01)
G06V 10/30(2022.01)
G06N 3/0455(2023.01)
G06N 3/08(2023.01)

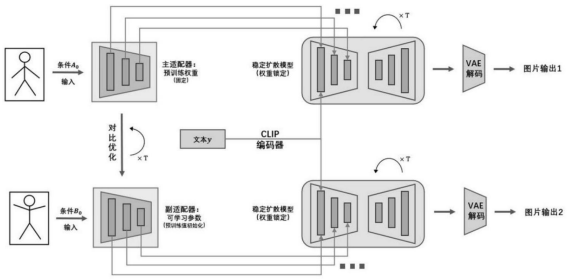
权利要求书3页 说明书14页 附图1页

(54)发明名称

一种基于适配网络增强扩散模型的人体姿态场景恢复方法

(57)摘要

本发明公开了一种基于适配网络增强扩散模型的人体姿态场景恢复方法,属于计算机视觉的图像生成领域,该方法首先每一组训练用的数据表示为一个五元组$(A^{(N)},B^{(N)},y,Z_A,Z_B)$,并用点阵$A^{(N)}$,$B^{(N)}$分别制作灰度图$A_0$,$B_0$。其次由扩散模型得到扩散模型中主副适配器的差异,并根据差异计算出损失函数,并计算结余损失。最后根据损失函数和结余损失函数得到全局优化函数,对输出的人物姿态图进行优化。本发明消除了通过试探方法训练寻找损失函数间权值时所需的计算开销,实现了预训练模型功能细分的训练方式,使得输出的人物姿态图更为准确稳定。



CN 117392675 A

1.一种基于适配网络增强扩散模型的人体姿态场景恢复方法,其特征在于,包括以下步骤:

步骤1.每一组训练用的数据表示为一个五元组($A^{(N)}$,$B^{(N)}$,y,Z_A,Z_B),其中y为文本描述标注,$A^{(N)}$,$B^{(N)}$是两个元素个数均为$\mathcal{N}$的点阵,$\boldsymbol{Z_A}, \boldsymbol{Z_B} \in \mathbb{R}^{3\times w\times h}$为图像数据,其中数字3意味着图像按照RGB格式存储,w,h分别表示图像的宽度和高度;用$\Psi_\delta(\bullet)$表示输入图像产生图像中人物关节基点坐标点点集的模型,以δ为权重;根据图像Z_A,Z_B分别得到的点阵$A^{(N)}$,$B^{(N)}$;

步骤2.用点阵$A^{(N)}$,$B^{(N)}$分别制作灰度图A_0,B_0;

步骤3.由扩散模型得到扩散模型中主副适配器的差异,并根据差异计算出损失函数,并计算结余损失;

步骤4.根据损失函数和结余损失函数得到全局优化函数,对输出的人物姿态图进行优化。

2.根据权利要求1所述的基于适配网络增强扩散模型的人体姿态场景恢复方法,其特征在于,在步骤1中,得到所述点阵$A^{(N)}$,$B^{(N)}$的具体过程为:

1.1.文本描述标注y∈{y_A,y_B},y_A,y_B分别表示图像Z_A,Z_B中图像内容对应的英文文本描述;

1.2.通过模型$\Psi_\delta(\bullet)$计算输入图像Y中的关节基点点集的过程描述为:

$$\Psi_\delta(\boldsymbol{Y}) = [\boldsymbol{y}_1^T \quad \boldsymbol{y}_2^T \quad \dots \quad \boldsymbol{y}_N^T]^T \tag{1}$$

用$\boldsymbol{y}_i = (x_i, y_i) \in \mathbb{R}^2$表示第i个基点在w×h大小的坐标系中的坐标位置,并用$y_{A,i}$表示图像A中第i个基点的纵坐标位置,$y_{B,i}$同理,类似定义$x_{A,i}$,$x_{B,i}$;通过这个模型得到一个有$\mathcal{N}$个点组成的点阵;使用如下定义的归一化函数对点阵中的点进行归一化,即压缩,对某个点$\boldsymbol{y}_i \in \mathbb{R}^2$的归一化表示为:

$$g_b(\boldsymbol{y}_i) = \Psi_\delta(\boldsymbol{Y}) \begin{bmatrix} 1/b_w & 0 \\ 0 & 1/b_h \end{bmatrix} (\boldsymbol{y}_i - \boldsymbol{b}_c) \tag{2}$$

其中b_w=w,b_h=h,$\boldsymbol{b}_c \in \mathbb{R}^2$为可调偏置;

对整个点阵,压缩过程表示为:

$$g_b\left(\Psi_\delta(\boldsymbol{Y})\right) = g_b([\boldsymbol{y}_1^T \quad \boldsymbol{y}_2^T \quad \dots \quad \boldsymbol{y}_N^T]^T) = [g_b(\boldsymbol{y}_1^T) \quad g_b(\boldsymbol{y}_2^T) \quad \dots \quad g_b(\boldsymbol{y}_N^T)]^T \tag{3}$$

结合式(1),(3),用复合函数的形式表示为:

$$\Psi_\delta \circ g_b(\boldsymbol{Y}) = [g_b(\boldsymbol{y}_1^T) \quad g_b(\boldsymbol{y}_2^T) \quad \dots \quad g_b(\boldsymbol{y}_N^T)]^T \tag{4}$$

则点阵$\boldsymbol{A}^{(\mathcal{N})}, \boldsymbol{B}^{(\mathcal{N})}$的计算方式如下:

$$\boldsymbol{A}^{(N)} = \Psi_\delta \circ g_b(\boldsymbol{Z}_A) = \left[y_{A,1}^T \quad y_{A,2}^T \quad \dots \quad y_{A,\mathcal{N}}^T\right]^T \tag{5}$$

$$\boldsymbol{B}^{(N)} = \Psi_\delta \circ g_b(\boldsymbol{Z}_B) = \left[y_{B,1}^T \quad y_{B,2}^T \quad \dots \quad y_{B,\mathcal{N}}^T\right]^T \tag{6}$$

其中$\boldsymbol{y}_{A,i} = \left[x_{A,i}, y_{A,i}\right]^T \in \mathbb{R}^2, \boldsymbol{y}_{B,i} = \left[x_{B,i}, y_{B,i}\right]^T \in \mathbb{R}^2$。

3.根据权利要求2所述的基于适配网络增强扩散模型的人体姿态场景恢复方法,其特征在于,步骤2具体过程如下:

将压缩后的点阵等比扩展到[0,u]×[0,u],u=min{w,h};

将点阵中出现过的点对应的坐标位置的灰度值设为0,其余位置设为255,分别获得两

张灰度图A_0,B_0。

4.根据权利要求3所述的基于适配网络增强扩散模型的人体姿态场景恢复方法,其特征在于,步骤2还包括,分别将原始图像Z_A,Z_B传入文本转换模型,获得的文本的向量编码为$\tau(y)$,每一条数据的五元组变为:(A_0,B_0,$\tau(y)$,Z_A,Z_B),同时设置可调参数$T \in \mathbb{Z}$,表示扩散模型的扩散总步长;分别取$y=y_A$或者$y=y_B$,得到两条数据;

使用原始图像制作出的灰度图A_0,B_0为最终训练时所需的图像数据;将实际训练时用到的数据集记TRAIN,对所有的灰度图组合A_0,B_0:$\forall(\boldsymbol{A_0},\boldsymbol{B_0},\boldsymbol{y_A},\boldsymbol{Z_A},\boldsymbol{Z_B}) \in \boldsymbol{TRAIN}$,($B_0$,$A_0$,$y_B$,$Z_B$,$Z_A$) $\in$TRAIN。

5.根据权利要求4所述的基于适配网络增强扩散模型的人体姿态场景恢复方法,其特征在于,步骤3具体过程如下:

3.1.将扩散模型的推理过程描述为:

$$\bar{\boldsymbol{Z}}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\mathbf{Z}}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\mathbf{Z}}_t,t)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\mathbf{Z}}_t,t) \quad (7)$$

上式中$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$,且$\alpha_t \in [0,1]$为扩散模型中的可调参数;

扩散模型中,时间步骤t遍历T,T-1,...,2,1,共T个等式,式中ϵ_θ为扩散模型中预训练的U-Net模型,以θ为权重;$\bar{\mathbf{Z}}_t$表示去噪过程在第t步采样得到生成图像在隐空间中的表示,且在t=T时取$\bar{\mathbf{Z}}_T=\epsilon\sim\mathcal{N}(0,\boldsymbol{I})$;$\mathcal{N}(0,\boldsymbol{I})$表示服从各维度均值均为0,协方差矩阵为单位矩阵I,即一个高维正态分布;

3.2.将扩散模型中的隐空间图像表示解码成真实图片,使用变分自编码器的解码器,记为De(•),t时的真实图像Z_t表示为:

$$\boldsymbol{Z}_t=\boldsymbol{De}(\bar{\mathbf{Z}}_t) \quad (8)$$

在使用适配器的扩散模型的去噪过程中,隐空间中的图像表示为:

$$\bar{\boldsymbol{Z}}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\mathbf{Z}}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right) \quad (9)$$

其中X_0表示图像条件输入,y表示文本输入,F_Φ表示以Φ为权重的适配器模型;

在上述扩散模型中,分别将灰度图A_0传入主适配器$\boldsymbol{F}_{\Phi_1}$获得向量$\boldsymbol{F}_{\Phi_1}(\boldsymbol{A_0})$,将灰度图$B_0$传入副适配器$\boldsymbol{F}_{\Phi_2}$获得向量$\boldsymbol{F}_{\Phi_2}(\boldsymbol{B_0})$,分别按照T2I-Adapter模型中的方法注入到扩散模型中,Φ_1,Φ_2分别表示适配器模型使用的权重;

3.3.扩散模型使用U-Net作为去噪网络,并使用对比语言图像预训练CLIP作为将文本内容转化为向量的编码器,使用免类别指引;

将主适配器所在生成路径记为线路1,副适配器所在生成路径记为线路2;对于路线1,将在时间步骤为t时扩散模型采样得到的图像记为$\boldsymbol{Z}_{A,t}=\boldsymbol{De}(\bar{Z}_{A,t})$,类似定义$Z_{B,t}$;

路线1与路线2中主副适配器的差异表示为:

$$\left|\left|\nabla log\, p(\bar{\boldsymbol{Z}}_{A,t})-\nabla log\, p(\bar{\boldsymbol{Z}}_{B,t})\right|\right|_2^2 \quad (10)$$

结合Tweedie公式有:

$$\frac{1}{1-\bar{\alpha}_t}\cdot\left\|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right\|_2^2 \quad (11)$$

对上式求期望得损失函数：

$$\mathcal{L}_{diffusion}$$

$$=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left\|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(y),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right\|_2^2\right] \quad (12)$$

3.4.设置结余损失函数，从损失函数上限制姿态，结余损失函数为：

$$\mathcal{L}_{const}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right\|_2^2\right] \quad (13)。$$

6.根据权利要求5所述的基于适配网络增强扩散模型的人体姿态场景恢复方法，其特征在于，步骤4所述全局优化函数如下：

$$\mathcal{L}_{\Phi_2}=2\cdot\mathbb{E}_{t\sim U[1,T]}\left[\frac{\alpha_t}{1-\alpha_t}\right]\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g_b(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g_b(\boldsymbol{B}_0)\right\|_2^2\right]$$

$$+\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right\|_2^2\right] \quad (14)。$$

7.根据权利要求6所述的基于适配网络增强扩散模型的人体姿态场景恢复方法，其特征在于，步骤4中还包括，全局损失收敛理理论值为：

$$\Delta\mathcal{L}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B,),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g(\boldsymbol{B}_0)-\Psi_\delta\circ g(\boldsymbol{A}_0)\right\|_2^2\right] \quad (15)$$

判断模型是否充分训练。

一种基于适配网络增强扩散模型的人体姿态场景恢复方法

技术领域

[0001] 本发明属于计算机视觉的图像生成领域,具体涉及一种基于适配网络增强扩散模型的人体姿态场景恢复方法。

背景技术

[0002] 计算机视觉领域的图像生成任务最早由生成对抗网络(GAN)完成,但近年来其逐渐被效果不断提高稳定扩散模型(Stable-Diffusion)取代,能够通过输入的文本信息产生与之对应的图片。但以上两种框架下的模型都还无法根据给定输入生成特定场景、特定对象、以及特定轮廓中下人物或物体。传统模式下直接生成人像或物像的模型大都专注于如何通过准确理解文本提示,从而在更精确的文本提示下实现图像操作(例如对输入文本生成图像以及按照文本要求对图像进行修改)。

[0003] 稳定生成图像的关键就在于通过将需要表达的要求信息通过合理形式传递给机器,并使其得到编码产生输出。究其根本,这是因为人类语言描述在机器视角下存在非常大的不确定因素,实际存在的模糊语义远超想象。在ControlNet架构以及T2I-Adapter架构提供了使用诸如语义分割、关键姿态识别等小模型提供扩散模型以图像约束并给出了将额外约束条件注入到扩散模型中的适配器(Adapter)模型;但此类模型在注入时会与文本引导产生混淆。具体表现为:在使用相同文字引导时使用同一小模型给出不同图像约束时,最终生成内容的场景乃至风格模式会产生极大的漂移,表现出低鲁棒性。

发明内容

[0004] 针对上述问题,本发明提出了一种基于适配网络增强扩散模型的人体姿态场景恢复方法,该方法基于T2I-Adapter的架构进行主、副适配器(Primary-Adapter and Secondary-Adapter)的联合训练方法,同时推导并简化了训练的优化目标,构建出基于微调适配器的模型,实现了通过适配器嵌入一个诸如关键姿态识别、图像语义分割的小模型,根据小模型分析得到的信息对图像根据文本的内容区分和还原出小模型检测信息不变的背景场景(例如对于关键姿态识别的小模型,针对不同的关键姿态恢复出文本指引的背景场景,同时保持关键姿态特征中的对象不变)。该发明旨在通过微调模型对锁定的大权重进行操控,以更加可靠和可控的小模型采样高维像素分布,驱动更高要求的图像生成任务。

[0005] 本发明主要通过引入副适配器(Secondary-Adapter,简写为S-ad),在固定stable-diffusion权重和预训练主适配器(Primary-Adapter,简写为P-ad)的权重条件下,微调副适配器,从而达到副适配器针对主适配器输出稳定的效果。同时,经过微调的副适配器拥有针对文本的稳定性;配合Google-Dreambooth在生成方面的特征提取模型,能够轻松实现:任意指定场景(文本描述)、任意指定人物(图片特征输入)、任意指定姿势(关键姿态)的图像生成。

[0006] 本发明中先构建出两个局部优化目标,再通过变换将两个目标映射到相同尺度进行联合训练;同时对该方在搭配Stable-Diffusion的主、副适配器模型中进行推广,使其能

够适应使用不同的小模型对Stable-Diffusion进行驱动的功能。

[0007] 本发明提供如下技术方案：

[0008] 步骤1.每一组训练用的数据表示为一个五元组$(A^{(\mathcal{N})},B^{(\mathcal{N})},y,Z_A,Z_B)$，其中y为文本描述标注，$A^{(\mathcal{N})},B^{(\mathcal{N})}$是两个元素个数均为$\mathcal{N}$的点阵，$\boldsymbol{Z_A},\boldsymbol{Z_B}\in\mathbb{R}^{3\times w\times h}$为图像数据，其中数字3意味着图像按照RGB格式存储，w,h分别表示图像的宽度和高度；用$\Psi_\delta(\bullet)$表示输入图像产生图像中人物关节基点坐标点点集的模型，以δ为权重。

[0009] 作为补充说明，文本描述标注$y\in\{y_A,y_B\}$，这里y_A,y_B分别表示图像Z_A,Z_B中图像内容对应的英文文本描述(因此这里实际上会产生两个五元组)。

[0010] 通过模型$\Psi_\delta(\bullet)$计算输入图像Y中的关节基点点集的过程描述为：

[0011] $$\Psi_\delta(\boldsymbol{Y})=[\boldsymbol{y}_1^T\quad\boldsymbol{y}_2^T\quad\dots\quad\boldsymbol{y}_{\mathcal{N}}^T]^T \tag{1}$$

[0012] 之后用$\boldsymbol{y}_i=(x_i,y_i)\in\mathbb{R}^2$表示第i个基点在w×h大小的坐标系中的坐标位置，并用$y_{A,i}$表示图像A中第i个基点的纵坐标位置，$y_{B,i}$同理，类似定义$x_{A,i},x_{B,i}$；通过这个模型得到一个有$\mathcal{N}$个点组成的点阵(点集)。使用如下定义的归一化函数对点阵中的点进行归一化(即压缩)，具体来说，对某个点$\boldsymbol{y}_i\in\mathbb{R}^2$的归一化表示为：

[0013] $$g_{\boldsymbol{b}}(\boldsymbol{y}_i)=\Psi_\delta(\boldsymbol{Y})\begin{bmatrix}1/b_w & 0\\ 0 & 1/b_h\end{bmatrix}(\boldsymbol{y}_i-\boldsymbol{b}_c) \tag{2}$$

[0014] 这里$b_w=w,b_h=h,\boldsymbol{b}_c\in R^2$为可调偏置(超参数之一)。

[0015] 对整个点阵，归一化(压缩)过程表示为：

[0016] $$g_{\boldsymbol{b}}\left(\Psi_\delta(\boldsymbol{Y})\right)=g_{\boldsymbol{b}}([\boldsymbol{y}_1^T\quad\boldsymbol{y}_2^T\quad\dots\quad\boldsymbol{y}_{\mathcal{N}}^T]^T)=[g_{\boldsymbol{b}}(\boldsymbol{y}_1^T)\quad g_{\boldsymbol{b}}(\boldsymbol{y}_2^T)\quad\dots\quad g_{\boldsymbol{b}}(\boldsymbol{y}_{\mathcal{N}}^T)]^T \tag{3}$$

[0017] 结合式(1)，(3)，用复合函数的形式可以表示为：

[0018] $$\Psi_\delta\circ g_{\boldsymbol{b}}(\boldsymbol{Y})=[g_{\boldsymbol{b}}(\boldsymbol{y}_1^T)\quad g_{\boldsymbol{b}}(\boldsymbol{y}_2^T)\quad\dots\quad g_{\boldsymbol{b}}(\boldsymbol{y}_{\mathcal{N}}^T)]^T \tag{4}$$

[0019] 根据以上定义，数据五元组中两个元素个数均为$\mathcal{N}$的点阵$A^{(\mathcal{N})},B^{(\mathcal{N})}$根据如下方式计算得到：

[0020] $$\boldsymbol{A}^{(\mathcal{N})}=\Psi_\delta\circ g_b(\boldsymbol{Z}_A)=\left[\boldsymbol{y}_{\boldsymbol{A},1}^T\quad\boldsymbol{y}_{\boldsymbol{A},2}^T\quad\dots\quad\boldsymbol{y}_{\boldsymbol{A},\mathcal{N}}^T\right]^T \tag{5}$$

[0021] $$\boldsymbol{B}^{(\mathcal{N})}=\Psi_\delta\circ g_{\boldsymbol{b}}(\boldsymbol{Z}_{\boldsymbol{B}})=\left[\boldsymbol{y}_{\boldsymbol{B},1}^T\quad\boldsymbol{y}_{\boldsymbol{B},2}^T\quad\dots\quad\boldsymbol{y}_{\boldsymbol{B},\mathcal{N}}^T\right]^T \tag{6}$$

[0022] 其中$\boldsymbol{y}_{\boldsymbol{A},i}=\left[x_{\boldsymbol{A},i},y_{\boldsymbol{A},i}\right]^T\in\mathbb{R}^2,\boldsymbol{y}_{\boldsymbol{B},i}=\left[x_{B,i},y_{B,i}\right]^T\in\mathbb{R}^2$。

[0023] 根据图像Z_A,Z_B分别得到的点阵$\boldsymbol{A}^{(\mathcal{N})},\boldsymbol{B}^{(\mathcal{N})}$，在后续的步骤中分别渲染成灰度图像。

[0024] 步骤2.将步骤1.中获得的两个点阵$\boldsymbol{A}^{(\mathcal{N})},\boldsymbol{B}^{(\mathcal{N})}$分别制作灰度图$A_0,B_0$：将压缩后的点阵等比扩展到[0,u]×[0,u]，这里u=min{w,h}；另外，将点阵中出现过的点对应的坐标位置的灰度值设为0，其余位置设为255，分别获得两张灰度图A_0,B_0。

[0025] 分别将原始图像Z_A,Z_B传入文本转换模型(将文本内容转化为向量编码)，获得的文本的向量编码为τ(y)，每一条数据的五元组变为：

[0026] $$(A_0,B_0,\tau(y),Z_A,Z_B) \tag{7}$$

[0027] 同时设置可调参数$T\in\mathbb{Z}$，表示扩散模型的扩散总步长；这里分别取y=y_A或者y=y_B，可得到两条数据。

[0028] 实际上,在这个过程中使用原始图像制作出的灰度图A_0,B_0为最终训练时所需的图像数据。将实际训练时用到的数据集记TRAIN,对所有的灰度图组合A_0,B_0:

[0029] $$\forall(\boldsymbol{A}_0,\boldsymbol{B}_0,\boldsymbol{y}_A,\boldsymbol{Z}_A,\boldsymbol{Z}_B)\in\boldsymbol{TRAIN},(\boldsymbol{B}_0,\boldsymbol{A}_0,\boldsymbol{y}_B,\boldsymbol{Z}_B,\boldsymbol{Z}_A)\in\boldsymbol{TRAIN} \quad (8)$$

[0030] 步骤3.将扩散模型的去噪过程(采样过程,或推理过程)描述为:

[0031] $$\bar{\boldsymbol{Z}}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\mathbf{Z}}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\mathbf{Z}}_t,t)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\mathbf{Z}}_t,t) \quad (9)$$

[0032] 上式中$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$,且$\alpha_t\in[0,1]$为扩散模型中的可调参数。

[0033] 在整个扩散模型中,t遍历T,T-1,...,2,1(扩散模型中的可调参数为α_t,$\sigma_t\in[0,1]$),共T个等式,按照t的取值分别叫做“第t个等式”;式中ϵ_θ为扩散模型中预训练的U-Net模型,以θ为权重。$\overline{\boldsymbol{Z}_t}$表示去噪过程在第t步采样得到生成图像在隐空间中的表示,且在t=T时取$\bar{\mathbf{Z}}_T=\epsilon\sim\mathcal{N}(0,\boldsymbol{I})$。

[0034] 进一步的,这里$\mathcal{N}(0,\boldsymbol{I})$表示服从各维度均值均为0,协方差矩阵为单位矩阵I,即一个高维正态分布;本发明中所有I的维数相同。

[0035] 将隐空间图像表示解码成真实图片,需要使用变分自编码器(VAE)的解码器,这个模型记为De(•),时间t下的真实图像Z_t表示为:

[0036] $$\boldsymbol{Z}_t=\boldsymbol{De}(\bar{\mathbf{Z}}_t) \quad (10)$$

[0037] 在使用适配器的扩散模型的去噪过程中,隐空间中的图像表示为:

[0038] $$\bar{\mathbf{Z}}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\mathbf{Z}}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right) \quad (11)$$

[0039] 其中X_0表示图像条件输入,y表示文本输入,F_Φ表示以Φ为权重的适配器模型。

[0040] 在上述扩散模型中,分别将步骤2.中得到的灰度图A_0传入主适配器(Primary-Adapter)$\boldsymbol{F}_{\Phi_1}$获得向量$\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)$,将灰度图$B_0$传入副适配器(Secondary-Adapter)$\boldsymbol{F}_{\Phi_2}$获得向量$\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)$,分别按照T2I-Adapter模型中的方法注入到扩散模型中,这里Φ_1,Φ_2分别表示适配器模型使用的权重。

[0041] 另外,本发明中使用到的扩散模型使用U-Net作为去噪网络,并使用对比语言图像预训练CLIP作为将文本内容转化为向量的编码器,使用免类别指引,即Classifier-Free Guidance,在扩散模型的每一个时间步(timestep)中不断地将这两个向量分别注入到Denoising U-Net中,固定Stable-Diffusion的权重θ和主适配器(Primary-Adapter)的权重Φ_1并设置副适配器(Secondary-Adapter)权重Φ_2为可学习参数,两条扩散生成路径计算噪声的预测损失$\mathcal{L}_{\Phi_2}$,推导过程如下:

[0042] 方便起见,将主适配器所在生成路径记为线路1,副适配器所在生成路径记为线路2。在训练主适配器$\boldsymbol{F}_{\Phi_1}$时使用的损失为:

[0043] $$\mathcal{L}_{\Phi_1}=\mathbb{E}_{(Z_A,A_0,y),t\sim U[1,T],\epsilon\sim N(0,I)}\left[w(t)\left\|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{a,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)\right\|_2^2\right] \quad (12)$$

[0044] 这里：

[0045] $$w(t)=\frac{1-\alpha_t}{2\alpha_t(1-\bar{\alpha}_t)} \quad (13)$$

[0046] 并用$\mathbb{E}$表示数学期望，$t\sim\mathcal{U}[1,T]$表示时间步长t服从[1,T]上的均匀分布，$\in\sim(0,I)$表示随机噪声$\in$服从随机正态分布，$\in_\theta$表示扩散模型中以θ为权重的U-Net网计算出的噪声预测值。

[0047] 对于路线1，将在时间步(timestep)为t时扩散模型采样得到的图像记为$\boldsymbol{Z}_{A,t}=\boldsymbol{De}(\bar{Z}_{A,t})$，类似定义$Z_{B,t}$。

[0048] 为了实现通过关节基点控制人像姿态的目标，需要保证1，2线路中采样预测噪声得到的图像，除关节基点不同造成差异分布像素外的其他像素分布尽量相同。由于扩散模型中使用梯度量化表示扩散方向的得分，因此在某个时间步(timestep)下，路线1与路线2中扩散噪声的差异可以表示为：

[0049] $$\left|\left|\nabla log\, p(\bar{\boldsymbol{Z}}_{A,t})-\nabla log\, p(\bar{\boldsymbol{Z}}_{B,t})\right|\right|_2^2 \quad (14)$$

[0050] 这里p(•)表示求概率分布，∇表示求梯度，结合Tweedie公式可知：

[0051] $$\nabla\log p(\bar{\boldsymbol{Z}}_{A,t})=-\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_0\approx-\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right) \quad (15)$$

[0052] 以及：

[0053] $$\nabla\log p(\bar{\boldsymbol{Z}}_{B,t})=-\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_0\approx-\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right) \quad (16)$$

[0054] 这量化表示了模型内部的扩散方向，而分别经过两个适配器注入向量后的扩散模型，在时间步t下扩散方向上的差异可以被表示为：

[0055] $$\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right|\right|_2^2 \quad (17)$$

[0056] 因此，根据式15，16设计针对线路1与线路2扩散方向的优化，设置损失函数：

[0057] $$\mathcal{L}_{diffusion}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(y),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right|\right|_2^2\right] \quad (18)$$

[0058] 另一方面，考虑到实际上线路1与线路2中主、副适配器注入后的扩散结果差异应当被限制在关节基点标注出的姿态上，故还需要从损失函数上限制姿态，需保证：

[0059] $$\forall t\in[1,T],\quad\left|\left|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2=\left|\left|\Psi_\delta\circ g(\boldsymbol{B}_0)-\Psi_\delta\circ g(\boldsymbol{A}_0)\right|\right|_2^2 \quad (19)$$

[0060] 上式的右边实际上是一个常数。为满足式(19)限制的条件，可选择优化：

[0061] $$\mathcal{L}_{const}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\left|\left|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2-\left|\left|\Psi_\delta\circ g(\boldsymbol{B}_0)-\Psi_\delta\circ g(\boldsymbol{A}_0)\right|\right|_2^2\right] \quad (20)$$

[0062] 可以舍去其中的常数项，得到结余损失函数：

[0063] $$\mathcal{L}_{const}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_{\delta}\circ g(\boldsymbol{Z}_{B,t})-\Psi_{\delta}\circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2\right] \quad (21)$$

[0064] 至此,通过式(18),(21)得到了两个用于局部优化的损失函数。

[0065] 步骤4.在该步进行全局优化函数的确定。针对式(18)与(21),由于两个局部损失函数的尺度(scale)不同,不能直接相加。针对这个问题,传统方法大多将其中一个配上系数(这个系数在训练过程中试探出合适的值)后将两个或更多用于局部优化的函数相加,在这一步中使用一种全新的办法将两个用于局部优化的损失函数直接转换到同一尺度中,对输出的人物姿态图进行优化。

[0066] 由于优化主适配器得到的模型使稳定扩散(Stable-Diffusion)模型在固定权重中采样时可以稳定的让像素分布到关节基点位置(标定了人像的姿态),因此相当于在训练适配器时可以看作完成了一个存在限制条件的优化问题。

[0067] 设置全局优化函数为:

[0068] $$\mathcal{L}_{\Phi_2}=2\cdot\mathbb{E}_{t\sim U[1,T]}\left[\frac{\alpha_t}{1-\alpha_t}\right]\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_{\delta}\circ g_b(\boldsymbol{Z}_{B,t})-\Psi_{\delta}\circ g_b(\boldsymbol{B}_0)\right|\right|_2^2\right]$$

$$+\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_{\delta}\circ g(\boldsymbol{Z}_{B,t})-\Psi_{\delta}\circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2\right] \quad (22)$$

[0069] 作为推导模型正确性的检验,可以检查迭代训练(次数充分大)结束后的全局损失是否收敛到了理论值:

[0070] $$\Delta\mathcal{L}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B,),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_{\delta}\circ g(\boldsymbol{B}_0)-\Psi_{\delta}\circ g(\boldsymbol{A}_0)\right|\right|_2^2\right] \quad (23)$$

[0071] 本发明具有以下优势:本发明将两个局部优化函数变换到同一尺度,消除了通过试探方法训练寻找损失函数间权值时所需的计算开销;将直接优化损失函数转化为通过优化上确界(最小上界),提出了一种通过比较扩散模型中各步下生成效果,以实现预训练模型功能细分的训练方式,使得输出的人物姿态图更为准确稳定。

附图说明

[0072] 图1为本发明的训练流程图;

[0073] 图2为单适配器的向量注入方式;

[0074] 图3为本发明结果示意图。

具体实施方式

[0075] 以下结合附图,对本发明做深入解释。

[0076] 一种基于适配网络增强扩散模型的人体姿态场景恢复方法,如图1所示,包括以下具体步骤:

[0077] 步骤1.每一组训练用的数据表示为一个五元组($A^{(N)}$,$B^{(N)}$,y,Z_A,Z_B),其中y为文本描述标注,$A^{(N)}$,$B^{(N)}$是两个元素个数均为$\mathcal{N}$的点阵,$\boldsymbol{Z}_A,\boldsymbol{Z}_B\in\mathbb{R}^{3\times w\times h}$为图像数据,其中数字3意味着图像按照RGB格式存储,w,h分别表示图像的宽度和高度;用$\Psi_{\delta}(\bullet)$表示输入图像产生图像中人物关节基点坐标点点集的模型,以δ为权重。

[0078] 作为补充说明,文本描述标注$y\in\{y_A,y_B\}$,这里y_A,y_B分别表示图像Z_A,Z_B中图像内容对应的英文文本描述(因此这里实际上会产生两个五元组)。

[0079] 通过模型$\Psi_\delta(\cdot)$计算输入图像Y中的关节基点点集的过程描述为:

[0080] $$\Psi_\delta(\boldsymbol{Y})=[\boldsymbol{y}_1^T\quad\boldsymbol{y}_2^T\quad\cdots\quad\boldsymbol{y}_{\mathcal{N}}^T]^T \tag{1}$$

[0081] 之后用$\boldsymbol{y}_i=(x_i,y_i)\in\mathbb{R}^2$表示第i个基点在w×h大小的坐标系中的坐标位置,并用$y_{A,i}$表示图像A中第i个基点的纵坐标位置,$y_{B,i}$同理,类似定义$x_{A,i},x_{B,i}$;通过这个模型得到一个有$\mathcal{N}$个点组成的点阵(点集)。使用如下定义的归一化函数对点阵中的点进行归一化(即压缩),具体来说,对某个点$\boldsymbol{y}_i\in\mathbb{R}^2$的归一化表示为:

[0082] $$g_{\boldsymbol{b}}(\boldsymbol{y}_i)=\Psi_\delta(\boldsymbol{Y})\begin{bmatrix}1/b_w & 0\\ 0 & 1/b_h\end{bmatrix}(\boldsymbol{y}_i-\boldsymbol{b}_c) \tag{2}$$

[0083] 这里$b_w=w,b_h=h,\boldsymbol{b}_c\in R^2$为可调为偏置(超参数之一)。

[0084] 对整个点阵,归一化(压缩)过程表示为:

[0085] $$g_{\boldsymbol{b}}\left(\Psi_\delta(\boldsymbol{Y})\right)=g_{\boldsymbol{b}}([\boldsymbol{y}_1^T\quad\boldsymbol{y}_2^T\quad\cdots\quad\boldsymbol{y}_{\mathcal{N}}^T]^T)=[g_{\boldsymbol{b}}(\boldsymbol{y}_1^T)\quad g_{\boldsymbol{b}}(\boldsymbol{y}_2^T)\quad\cdots\quad g_{\boldsymbol{b}}(\boldsymbol{y}_{\mathcal{N}}^T)]^T \tag{3}$$

[0086] 结合式(1),(3),用复合函数的形式可以表示为:

[0087] $$\Psi_\delta\circ g_{\boldsymbol{b}}(\boldsymbol{Y})=[g_{\boldsymbol{b}}(\boldsymbol{y}_1^T)\quad g_{\boldsymbol{b}}(\boldsymbol{y}_2^T)\quad\cdots\quad g_{\boldsymbol{b}}(\boldsymbol{y}_{\mathcal{N}}^T)]^T \tag{4}$$

[0088] 根据以上定义,数据五元组中两个元素个数均为$\mathcal{N}$的点阵$A^{(N)},B^{(N)}$根据如下方式计算得到:

[0089] $$\boldsymbol{A}^{(\mathcal{N})}=\Psi_\delta\circ g_{\boldsymbol{b}}(\boldsymbol{Z}_A)=\left[y_{\boldsymbol{A},1}^T\quad y_{\boldsymbol{A},2}^T\quad\cdots\quad y_{\boldsymbol{A},\mathcal{N}}^T\right]^T \tag{5}$$

[0090] $$\boldsymbol{B}^{(\mathcal{N})}=\Psi_\delta\circ g_{\boldsymbol{b}}(\boldsymbol{Z}_B)=\left[y_{\boldsymbol{B},1}^T\quad y_{\boldsymbol{B},2}^T\quad\cdots\quad y_{\boldsymbol{B},\mathcal{N}}^T\right]^T \tag{6}$$

[0091] 其中$\boldsymbol{y}_{\boldsymbol{A},i}=\left[x_{\boldsymbol{A},i},y_{\boldsymbol{A},i}\right]^T\in\mathbb{R}^2,\boldsymbol{y}_{\boldsymbol{B},i}=\left[x_{\boldsymbol{B},i},y_{\boldsymbol{B},i}\right]^T\in\mathbb{R}^2$。

[0092] 根据图像Z_A,Z_B分别得到的点阵$\boldsymbol{A}^{(\mathcal{N})},\boldsymbol{B}^{(\mathcal{N})}$,在后续的步骤中分别渲染成灰度图像。

[0093] 步骤2.将步骤1.中获得的两个点阵$\boldsymbol{A}^{(\mathcal{N})},\boldsymbol{B}^{(\mathcal{N})}$分别制作灰度图$A_0,B_0$:将压缩后的点阵等比扩展到[0,u]×[0,u],这里u=min{w,h};另外,将点阵中出现过的点对应的坐标位置的灰度值设为0,其余位置设为255,分别获得两张灰度图A_0,B_0。

[0094] 分别将原始图像Z_A,Z_B传入文本转换模型(将文本内容转化为向量编码),获得的文本的向量编码为τ(y),每一条数据的五元组变为:

[0095] $$(A_0,B_0,\tau(y),Z_A,Z_B) \tag{7}$$

[0096] 同时设置可调参数$\boldsymbol{T}\in\mathbb{Z}$,表示扩散模型的扩散总步长;这里分别取$y=y_A$或者$y=y_B$,可得到两条数据。

[0097] 实际上,在这个过程中使用原始图像制作出的灰度图A_0,B_0为最终训练时所需的图像数据。将实际训练时用到的数据集记TRAIN,对所有的灰度图组合A_0,B_0:

[0098] $$\forall(\boldsymbol{A}_0,\boldsymbol{B}_0,\boldsymbol{y}_A,\boldsymbol{Z}_A,\boldsymbol{Z}_B)\in\boldsymbol{TRAIN},(\boldsymbol{B}_0,\boldsymbol{A}_0,\boldsymbol{y}_B,\boldsymbol{Z}_B,\boldsymbol{Z}_A)\in\boldsymbol{TRAIN} \tag{8}$$

[0099] 步骤3.将扩散模型的去噪过程(采样过程,或推理过程)描述为:

[0100] $$\bar{\boldsymbol{Z}}_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\boldsymbol{Z}}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\boldsymbol{Z}}_t,t)}{\sqrt{\bar{\alpha}_t}}\right)+\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\boldsymbol{Z}}_t,t) \tag{9}$$

[0101] 上式中$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$,且$\alpha_t \in [0,1]$为扩散模型中的可调参数。

[0102] 在整个扩散模型中,t遍历T,T-1,...,2,1(扩散模型中的可调参数为$\alpha_t, \sigma_t \in [0,1]$),共T个等式,按照t的取值分别叫做“第t个等式”;式中ϵ_θ为扩散模型中预训练的U-Net模型,以θ为权重。$\overline{\boldsymbol{Z}_t}$表示去噪过程在第t步采样得到生成图像在隐空间中的表示,且在t=T时取$\bar{\mathbf{Z}}_T = \epsilon \sim \mathcal{N}(0, \boldsymbol{I})$。

[0103] 进一步的,这里$\mathcal{N}(0, \boldsymbol{I})$表示服从各维度均值均为0,协方差矩阵为单位矩阵I,即一个高维正态分布;本发明中所有I的维数相同

[0104] 将隐空间图像表示解码成真实图片,需要使用变分自编码器(VAE)的解码器,这个模型记为De(•),时间t下的真实图像Z_t表示为:

[0105] $$\boldsymbol{Z}_t = \boldsymbol{De}(\bar{\mathbf{Z}}_t) \tag{10}$$

[0106] 在使用适配器的扩散模型的去噪过程中,隐空间中的图像表示为:

[0107] $$\bar{\mathbf{Z}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\bar{\mathbf{Z}}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\overline{\boldsymbol{Z}_t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_\Phi(\boldsymbol{X}_0)\right) \tag{11}$$

[0108] 其中X_0表示图像条件输入,y表示文本输入,F_Φ表示以Φ为权重的适配器模型。

[0109] 在上述扩散模型中,分别将步骤2.中得到的灰度图A_0传入主适配器(Primary-Adapter)$\boldsymbol{F}_{\Phi_1}$获得向量$\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)$,将灰度图$B_0$传入副适配器(Secondary-Adapter)$\boldsymbol{F}_{\Phi_2}$获得向量$\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)$,分别按照T2I-Adapter模型中的方法注入到扩散模型中,这里Φ_1, Φ_2分别表示适配器模型使用的权重,单适配器的向量注入如图2所示。

[0110] 另外,本发明中使用到的扩散模型使用U-Net作为去噪网络,并使用对比语言图像预训练CLIP作为将文本内容转化为向量的编码器,使用免类别指引,即Classifier-Free Guidance,在扩散模型的每一个时间步(timesrep)中不断地将这两个向量分别注入到Denoising U-Net中,固定Stable-Diffusion的权重θ和主适配器(Primary-Adapter)的权重Φ_1并设置副适配器(Secondary-Adapter)权重Φ_2为可学习参数,两条扩散生成路径计算噪声的预测损失$\mathcal{L}_{\Phi_2}$,推导过程如下:

[0111] 方便起见,将附图中主适配器所在生成路径记为线路1,副适配器所在生成路径记为线路2。在训练主适配器$\boldsymbol{F}_{\Phi_1}$时使用的损失为:

[0112] $$\mathcal{L}_{\Phi_1} = \mathbb{E}_{(Z_A, A_0, y), t\sim U[1,T], \epsilon\sim N(0,I)}\left[w(t)\left\|\epsilon - \epsilon_\theta\left(\bar{\boldsymbol{Z}}_{a,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)\right\|_2^2\right] \tag{12}$$

[0113] 这里:

[0114] $$w(t) = \frac{1-\alpha_t}{2\alpha_t(1-\bar{\alpha}_t)} \tag{13}$$

[0115] 并用$\mathbb{E}$表示数学期望,$t \sim \mathcal{U}[1,T]$表示时间步长t服从[1,T]上的均匀分布,∈~(0,I)表示随机噪声∈服从随机正态分布,ϵ_θ表示扩散模型中以θ为权重的U-Net网计算出的噪声预测值。

[0116] 对于路线1,将在时间步(timestep)为t时扩散模型采样得到的图像记为

$\boldsymbol{Z}_{A,t} = \boldsymbol{De}(\bar{Z}_{A,t})$,类似定义$Z_{B,t}$。

[0117] 为了实现通过关节基点控制人像姿态的目标,需要保证1,2线路中采样预测噪声得到的图像,除关节基点不同造成差异分布像素外的其他像素分布尽量相同。由于扩散模型中使用梯度量化表示扩散方向的得分,因此在某个时间步(timestep)下,路线1与路线2中扩散噪声的差异可以表示为:

[0118] $$\left|\left|\nabla \log p(\bar{\boldsymbol{Z}}_{A,t}) - \nabla \log p(\bar{\boldsymbol{Z}}_{B,t})\right|\right|_2^2 \quad (14)$$

[0119] 这里p(•)表示求概率分布,∇表示求梯度,结合Tweedie公式可知:

[0120] $$\nabla \log p(\bar{\boldsymbol{Z}}_{A,t}) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_0 \approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right) \quad (15)$$

[0121] 以及:

[0122] $$\nabla \log p(\bar{\boldsymbol{Z}}_{B,t}) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_0 \approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right) \quad (16)$$

[0123] 这量化表示了模型内部的扩散方向,而分别经过两个适配器注入向量后的扩散模型,在时间步t下扩散方向上的差异可以被表示为:

[0124] $$\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right) - \epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right|\right|_2^2 \quad (17)$$

[0125] 因此,根据式15,16设计针对线路1与线路2扩散方向的优化,设置损失函数:

$$\mathcal{L}_{diffusion}$$

[0126] $$= \mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right) - \epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t}, t, \tau(y), \boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right|\right|_2^2\right] \quad (18)$$

[0127] 另一方面,考虑到实际上线路1与线路2中主、副适配器注入后的扩散结果差异应当被限制在关节基点标注出的姿态上,故还需要从损失函数上限制姿态,需保证:

[0128] $$\forall t \in [1,T], \quad \left|\left|\Psi_\delta \circ g(\boldsymbol{Z}_{B,t}) - \Psi_\delta \circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2 = \left|\left|\Psi_\delta \circ g(\boldsymbol{B}_0) - \Psi_\delta \circ g(\boldsymbol{A}_0)\right|\right|_2^2 \quad (19)$$

[0129] 上式的右边实际上是一个常数。为满足式(19)限制的条件,可选择优化:

$$\mathcal{L}_{const}$$

[0130] $$= \mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\left|\left|\Psi_\delta \circ g(\boldsymbol{Z}_{B,t}) - \Psi_\delta \circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2 - \left|\left|\Psi_\delta \circ g(\boldsymbol{B}_0) - \Psi_\delta \circ g(\boldsymbol{A}_0)\right|\right|_2^2\right] \quad (20)$$

[0131] 可以舍去其中的常数项,得到结余损失函数:

[0132] $$\mathcal{L}_{const} = \mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\left|\left|\Psi_\delta \circ g(\boldsymbol{Z}_{B,t}) - \Psi_\delta \circ g(\boldsymbol{Z}_{A,t})\right|\right|_2^2\right] \quad (21)$$

[0133] 至此,通过式(18),(21)得到了两个用于局部优化的损失函数

[0134] 另外,在基于公式(9)和(11)描述的扩散模型中,使用Tweedie公式推导式(15)和(16)的过程如下:

[0135] 对于高维高斯变量$\boldsymbol{z} \sim \mathcal{N}(z; \mu_z, \Sigma_z)$,其中$\mu_z$表示n维变量z的均值,$\Sigma_z$表示变量z中各维度变量间的协方差矩阵,Tweedie公式表明:

[0136] $$\mathbb{E}[\mu_z|\boldsymbol{z}] = \boldsymbol{z} + \Sigma_z \nabla_z \log p(\boldsymbol{z}) \quad (22)$$

[0137] 这里，$\mathbb{E}$表示数学期望，∇表示求梯度，p(·)表示求概率密度。

[0138] 在扩散模型中：

[0139] 条件概率$p(\boldsymbol{x}_t|\boldsymbol{x}_0) \sim N(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}, (1-\bar{\alpha})\boldsymbol{I})$ (23)

[0140] 代入Tweedie公式有：

[0141] $$\mathbb{E}[\mu_{x_t}|\boldsymbol{x}_t] = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla_{x_t} \log p(\boldsymbol{x}_t) \quad (24)$$

[0142] 这里可以将$\nabla_{x_t} \log p(\boldsymbol{x}_t)$简记为$\nabla \log p(\boldsymbol{x}_t)$

[0143] 而在一个预训练完成的扩散模型中，x_t的最佳生成来源于均值$\mu_{x_t} = \sqrt{\bar{\alpha}_t \boldsymbol{x}_0}$，故：

[0144] $$\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 = \boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t) \quad (25)$$

[0145] $$\Rightarrow \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}} \quad (26)$$

[0146] 与描述采样过程的式子比较得：

[0147] $$\nabla \log p(\boldsymbol{x}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_0 \approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta \quad (27)$$

[0148] 这里$\in_\theta$表示训练完备的去噪网络，在扩散模型中是卷积神经网络U-Net（扩散模型中先要完成噪声的预测，再按照去噪过程公式重新获得图像）。

[0149] 步骤4.在该步进行全局优化函数的确定。针对式(18)与(21)，由于两个局部损失函数的尺度(scale)不同，不能直接相加。针对这个问题，传统方法大多将其中一个配上系数(这个系数在训练过程中试探出合适的值)后将两个或更多用于局部优化的函数相加，在这一步中使用一种全新的办法将两个用于局部优化的损失函数直接转换到同一尺度中，对输出的人物姿态图进行优化。

[0150] 由于优化主适配器得到的模型使稳定扩散(Stable-Diffusion)模型在固定权重中采样时可以稳定的让像素分布到关节基点位置(标定了人像的姿态)，因此相当于在训练适配器时可以看作完成了一个存在限制条件的优化问题，以线路1为例：

[0151] $$\mathcal{L}_{\Phi_1} = \mathbb{E}_{(\bar{\boldsymbol{z}}_A, \boldsymbol{A}_0, \boldsymbol{y}), t\sim U[1,T], \epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[w(t)\left\|\epsilon - \epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(A_0)\right)\right\|_2^2\right] \quad (28)$$

$$s.t. \forall t \in [1,T] \quad \Psi_\delta \circ g_b(\boldsymbol{Z}_{A,t}) = \Psi_\delta \circ g_b(\boldsymbol{A}_0)$$

[0152] 这可以写为：

[0153] $$\mathcal{L}_{whole-\Phi_1} = \mathbb{E}_{(\bar{\boldsymbol{z}}_A, \boldsymbol{A}_0, \boldsymbol{y}), t\sim U[1,T], \epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[w(t)\left\|\epsilon - \epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t}, t, \tau(\boldsymbol{y}), \boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)\right\|_2^2\right]$$

$$+\lambda\mathbb{E}_{(\bar{\mathbf{z}}_B,\boldsymbol{A}_0,\boldsymbol{y}),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_\delta\circ g_b(\boldsymbol{Z}_{A,t})-\Psi_\delta\circ g_b(\boldsymbol{A}_0)\right|\right|_2^2\right]$$

[0154] $$=E_{(\boldsymbol{A}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}[w(t)\left|\left|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)\right|\right|_2^2$$

$$+\lambda\left|\left|\psi_\delta\circ g_b(\boldsymbol{Z}_{A,t})-\psi_\delta\circ g_b(\boldsymbol{A}_0)\right|\right|_2^2] \quad (29)$$

[0155] 这里λ＞0。

[0156] 式(12)中只有Φ_1是可学习参数,结合之前的分析有:

[0157] $$\Phi_1^*=\arg\min_{\Phi_1}\mathcal{L}_{\Phi_1}=\arg\min_{\Phi_1}\mathcal{L}_{\text{whole}-\Phi_1} \quad (30)$$

[0158] 因此使$\mathcal{L}_{\Phi_1}$和$\mathcal{L}_{whole-\Phi_1}$最优的$\Phi_1$相同,即:

[0159] $$\nabla_{\Phi_1}\mathcal{L}_{\Phi_1}=0\Leftrightarrow\nabla_{\Phi_1}\mathcal{L}_{whole-\Phi_1}=0 \quad (31)$$

[0160] 这表明:

[0161] $$\nabla_{\Phi_1}\mathbb{E}_{(\boldsymbol{Z}_A,\boldsymbol{A}_0,\boldsymbol{y}),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[w(t)\left|\left|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)\right|\right|_2^2\right]=0 \quad (32)$$

$$\Leftrightarrow\nabla_{\Phi_1}\mathbb{E}_{(\boldsymbol{Z}_A,\boldsymbol{A}_0,\boldsymbol{y}),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_\delta\circ g_b(\boldsymbol{Z}_{A,t})-\Psi_\delta\circ g_b(\boldsymbol{A}_0)\right|\right|_2^2\right]=0$$

[0162] 同理,对线路2有:

[0163] $$\nabla_{\Phi_2}\mathbb{E}_{(\boldsymbol{Z}_B,\boldsymbol{B}_0,\boldsymbol{y}),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[w(t)\left|\left|\epsilon-\epsilon_\theta\left(\boldsymbol{Z}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{B}_0)\right)\right|\right|_2^2\right]=0 \quad (33)$$

$$\Leftrightarrow\nabla_{\Phi_2}\mathbb{E}_{(\boldsymbol{Z}_B,\boldsymbol{B}_0,\boldsymbol{y}),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left|\left|\Psi_\delta\circ g_b(\bar{\boldsymbol{Z}}_{B,t})-\Psi_\delta\circ g_b(\boldsymbol{B}_0)\right|\right|_2^2\right]=0$$

[0164] 为简化表示,在之后的数学期望$\mathbb{E}$的下标中省去数据五元组(A_0,B_0,τ(y),Z_A,Z_B)

[0165] 对式(18),注意到:

[0166] $$\mathcal{L}_{diffusion}=\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)\right|\right|_2^2\right]$$

$$=\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left|\left|\left(\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)-\epsilon\right)-\left(\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)-\epsilon\right)\right|\right|_2^2\right]$$

$$\leq \mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{1-\bar{\alpha}_t}\cdot\left\|\left(\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A_0})\right)-\epsilon\right)\right\|_2^2\right.$$

$$\left.+\frac{1}{1-\bar{\alpha}_t}\cdot\left\|\left(\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B_0})\right)-\epsilon\right)\right\|_2^2\right]$$

$$=\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{(1-\bar{\alpha}_t)w(t)}\left(w(t)\cdot\left\|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A_0})\right)\right\|_2^2\right.\right.$$

[0167]

$$\left.\left.+w(t)\cdot\left\|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B_0})\right)\right\|_2^2\right)\right]$$

$$\leq \mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[\frac{1}{(1-\bar{\alpha}_t)w(t)}\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,I)}\left[w(t)\cdot\left\|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{A,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A_0})\right)\right\|_2^2\right.\right.$$

$$\left.\left.+w(t)\cdot\left\|\epsilon-\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B_0})\right)\right\|_2^2\right]\right] \tag{34}$$

[0168] 这里第一处放缩使用了L2线性空间中的三角不等式,第二处放缩使用了Cauchy不等式。

[0169] 考虑式(18)的上界中关于Φ2的梯度不为0的部分:

$$\sup_{\Phi_2}\mathcal{L}_{diffusion}$$

$$=\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\frac{1}{(1-\bar{\alpha}_t)w(t)}\right]\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[w(t)\cdot\left\|\epsilon-\right.\right.$$

[0170]

$$\left.\left.\epsilon_\theta\left(\bar{\boldsymbol{Z}}_{B,t},t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B_0})\right)\right\|_2^2\right]$$

$$\propto 2\cdot\mathbb{E}_{t\sim U[1,T]}\left[\frac{\alpha_t}{1-\alpha_t}\right]\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g_b(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g_b(\boldsymbol{B}_0)\right\|_2^2\right] \tag{35}$$

[0171] 结合式(18),(21),设置全局优化函数为:

$$\mathcal{L}_{\Phi_2}=2\cdot\mathbb{E}_{t\sim U[1,T]}\left[\frac{\alpha_t}{1-\alpha_t}\right]\cdot\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g_b(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g_b(\boldsymbol{B}_0)\right\|_2^2\right.$$

[0172]

$$\left.+\mathbb{E}_{t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right\|_2^2\right]\right. \tag{36}$$

[0173] 在这里将其称为对比误差,在之后计算该误差时,把前一项的系数记为Q,即:

[0174] $$Q=2\cdot\mathbb{E}_{t\sim U[1,T]}\left[\frac{\alpha_t}{1-\alpha_t}\right]=2\cdot\sum_{t=1}^{T}\frac{\alpha_t}{1-\alpha_t}\cdot t \tag{37}$$

[0175] 作为推导模型正确性的检验,可以检查迭代训练(次数充分大)结束后的全局损失是否收敛到了理论值:

[0176] $$\Delta\mathcal{L}=\mathbb{E}_{(\boldsymbol{A}_0,\boldsymbol{B}_0,\tau(\boldsymbol{y}),\boldsymbol{Z}_A,\boldsymbol{Z}_B,),t\sim U[1,T],\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\left\|\Psi_\delta\circ g(\boldsymbol{B}_0)-\Psi_\delta\circ g(\boldsymbol{A}_0)\right\|_2^2\right] \quad (38)$$

[0177] 实施例1：

[0178] 图1为整体模型的训练架构,以实现副适配器相对于主适配器的权重的稳定化,即扩散模型中副适配器相对于主适配器的稳定化微调训练方法。训练相对于使用者给定数据集上的样本稳定性时,需要提供关键姿态(Keypose)的图像形式输入数据。而实际上,由于训练建立在扩散模型中各步经过噪声预测和采样得到图像的关键姿态,因此并不需要保证获得关键姿态时来源图片中内容的相似性。据此分析,数据集中图像的具体内容对训练的结果影响不大,但为了避免最终训练出将文字描述内容在副适配器作用下强制扩散至不合理的关键姿态上的结果,使用图像描述(Image Caption)模型生成数据集中图像的描述。

[0179] 步骤1.选择公开数据集MPII-Human Pose Estimation,Cricket Shots Dataset以及Yogo Posture Dataset为原始图像数据集,旨在获得差异性较大的关键姿态数据,图片数量超过30000条,训练数据集制作时随机挑选其中的3000张图像。

[0180] 对图像数据集中的每一张原始图像Z_A(3000张图像中的1张),使用预训练的vit-apt2-image-captioning模型生成对图像中数据的描述,对图像内容进行标注。

[0181] 考虑到机器显存等计算中的实际问题,先设置下采样因子factor＝4,将图像的长宽都除以factor的值并取整作为下采样的目标形状,使用双线性插法(也可采用最近邻插值法、基于局部像素的重采样方法、基于4×4像素邻域的3次插值法或者基于8×8像素邻域的Lanczos插值法)进行下采样,用采样结果替代这个原始图像Z_A(即对所有原始图像进行等比缩小)。

[0182] 对每张图像Z_A,使用vit-gpt2-image-captioning(一种对输入图片进行文本描述的图像标注文本描述的模型)对随机挑选的1500张图片按照max_length＝30,beams＝5(max_length和beams均为vit-gpt2-image-captioning模型中的参数),获得图像对应的文本描述y_A。

[0183] 将按照上述方法处理后的图片作为输入,使用openpose模型获取对应的关键姿态(Keypose)的点阵输出为$A^{(\mathcal{N})}$,图像输出A_0。

[0184] 步骤2.根据步骤1.中的步骤,获得原始图像Z_A,Z_B,描述姿态的灰度图像A_0,B_0,以及对应描述图像的文本y_A,y_B。将最终用于训练的实际数据集记为TRAIN,有：

[0185] $$\forall(\boldsymbol{A}_0,\boldsymbol{B}_0,\boldsymbol{y}_A,\boldsymbol{Z}_A,\boldsymbol{Z}_B)\in\boldsymbol{TRAIN},\quad(\boldsymbol{B}_0,\boldsymbol{A}_0,\boldsymbol{y}_B,\boldsymbol{Z}_B,\boldsymbol{Z}_A)\in\boldsymbol{TRAIN} \quad (39)$$

[0186] 步骤3.每个五元组中的五个数据均为整个模型的输入,输入这些数据后按照ControlNet架构和T2I-Adapter提供的方法对扩散模型中的U-Net进行注入。对每一个时间步t,计算当前对比误差：

[0187] $$\mathcal{L}_t=Q\cdot\left\|\Psi_\delta\circ g_b(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g_b(\boldsymbol{B}_0)\right\|_2^2+\left\|\Psi_\delta\circ g(\boldsymbol{Z}_{B,t})-\Psi_\delta\circ g(\boldsymbol{Z}_{A,t})\right\|_2^2 \quad (40)$$

[0188] 这里：

[0189] $$\boldsymbol{Z}_{A,t}=\boldsymbol{De}\left(\frac{\epsilon-\sqrt{1-\overline{\alpha}_t}\epsilon_\theta\left(\boldsymbol{x}_t,t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_1}(\boldsymbol{A}_0)\right)}{\sqrt{\overline{\alpha}_t}}\right) \tag{41}$$

[0190] $$\boldsymbol{Z}_{B,t}=\boldsymbol{De}\left(\frac{\epsilon-\sqrt{1-\overline{\alpha}_t}\epsilon_\theta\left(\boldsymbol{x}_t,t,\tau(\boldsymbol{y}),\boldsymbol{F}_{\Phi_2}(\boldsymbol{B}_0)\right)}{\sqrt{\overline{\alpha}_t}}\right) \tag{42}$$

[0191] θ为稳定扩散模型的固定参数，Φ_1为固定主适配器参数，Φ_2为副适配器的可学习参数，函数De(•)表示对采样结果使用变分自编码(VAE)的解码器进行解码(decode)。

[0192] 步骤4.损失优化，重复步骤3的过程，取T＝50计算单步损失值，获得总损失为：

[0193] $$\mathcal{L}_{\Phi_2}=\sum_{t=1}^{T}\mathcal{L}_t \tag{43}$$

[0194] 进一步的，上述α_t(t＝1,2,...,T)为扩散模型中的可调参数，在这里取为初值为0.00002，末值为1.0的等差数列。

[0195] 实验结果如图3所示，实验表明，通过优选针对人体关节基点的检测模型，嵌入适配器，能够将前者的模型优越性迁移到扩散模型中来。相比于在ControlNet架构或T2I-Adapter下直接训练的模型，关节基点方面输入的细微变化会影响到模型对文本的反应上；副适配器(Secondary-Adapter)保留了主适配器(Primary-Adapter)根据文本产生的对像素的吸附能力的同时，防止了在去噪网络中注入向量时抹去文本编码的部分信息，从而做到了在给定相同文本的情况下，生成的图像仅在人体姿态上发生了改变。多适配器模式下，适配器的输入不再局限于人体姿势的定点，可以是语义分割、关键姿势等目标检测和识别模型，相应的L_2范数损失定义为相应模型中的损失计算。

[0196] 由于适配器在图像生成过程中只对注入时的生成形成了一定的干涉，整体扩散模型的生成能力依然由权重被锁定的稳定扩散模型(Stable-Diffusion)决定，副适配器同样产生对像素分布位置的干涉，但相对于文本稳定。

[0197] 原始的适配器，在ControlNet或T2I-Adapter架构下能以图像形式输入约束条件影响扩散模型的生成过程，使生成结果贴近约束条件图像；根据本发明训练出的副适配器权重，提高了约束条件注入时的稳定性，减少了约束条件注入对生成主线的影响，同时使模型通过文本恢复场景的能力得到极大的提高。

[0198] 本发明提出的方法，在4张NVIDIA-RTX4090显卡(单张显卡显存24GB，每张显卡对应拥有12核CPU与90GB内存，浮点算力：单精82.58TFLOPS/半精165.2Tensor TFLOPS)上花费约5小时，总损失最终收敛到理论值(误差0.04％)。通过对比不同框架与引导方式下的扩散模型以及本发明中的微调方式的生成效果，可以显示出本发明的优越性。

[0199] 实验结果如图3所示，使用“稳定扩散”模型(stable-diffusion model)，用对比语言图像预训练CLIP将输入文本编码后在隐空间中扩散，在输入“一位女孩，挥舞着左手”后，生成所示结果a，图片内容与文本有一定出入。使用ControlNet架构下的T2I-Adapter模型，输入“一位女孩，挥舞着左手”用对比语言图像预训练CLIP将输入文本编码后在隐空间中进

行扩散,同时适配器将输入的关键姿态进行编码投影到隐空间,分别与每一步的扩散结果相加,生成所示结果b,效果图的关键姿态与输入姿态比较吻合。依然使用ControlNet架构下的T2I-Adapter模型,使用通过本发明主副适配器联合训练方式微调后的主适配器,输入“一位女孩,挥舞着左手”用对比语言图像预训练CLIP将输入文本编码后在隐空间中进行扩散,同时适配器将输入的关键姿态进行编码投影到隐空间,分别与每一步的扩散结果相加,生成所示结果c,效果图的关键姿态与输入姿态基本完全吻合。

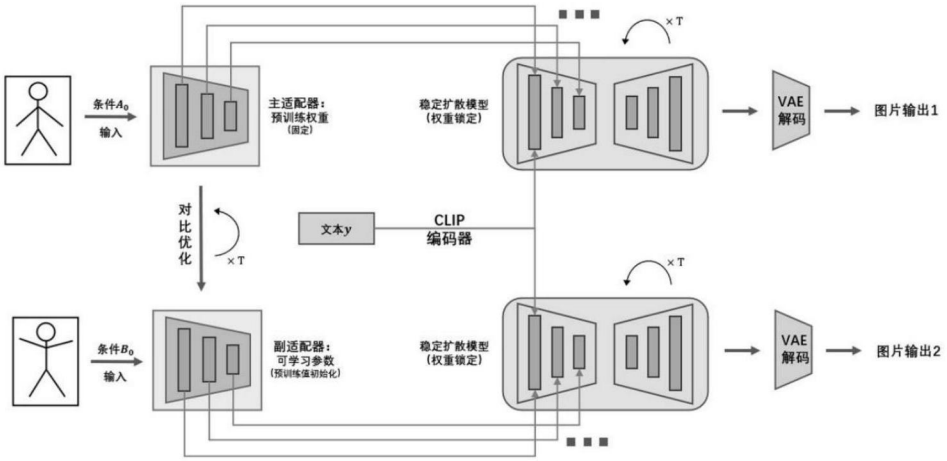

条件A_0
输入
主适配器：
预训练权重
(固定)
稳定扩散模型
(权重锁定)
×T
VAE
解码
图片输出1
对比优化
×T
文本y
CLIP
编码器
条件B_0
输入
副适配器：
可学习参数
(预训练值初始化)
稳定扩散模型
(权重锁定)
×T
VAE
解码
图片输出2

图1

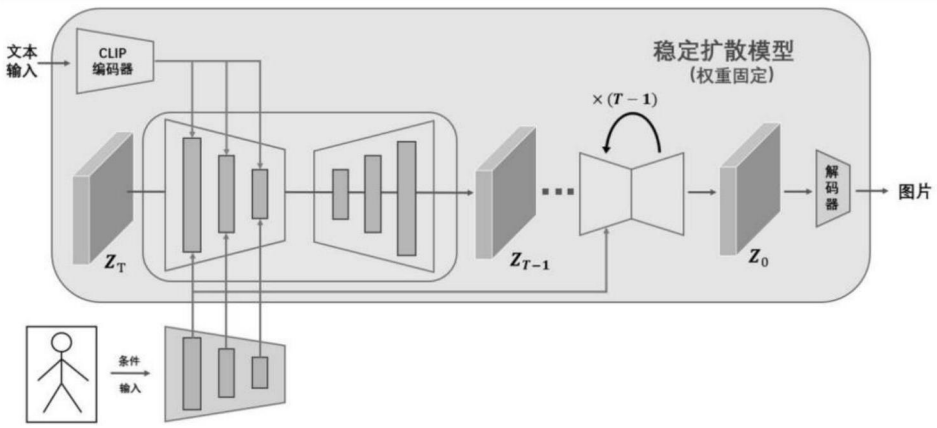

文本
输入
CLIP
编码器
稳定扩散模型
(权重固定)
×(T − 1)
Z_T
Z_{T-1}
Z_0
解码器
图片
条件
输入

图2


输入
a
b
c
+“一位女孩，挥舞着左手”

图3