基于最优传输理论的联合分布匹配方法及应用

曹杰彰[1)] 莫朗元[1)] 杜卿[1)] 国雍[1)] 赵沛霖[2)] 黄俊洲[2)] 谭明奎[1)]

[1)](华南理工大学 软件学院，广州 510006)

[2)](腾讯 AI Lab，深圳 518054)

摘 要 联合分布匹配问题是机器学习和计算机视觉领域的研究热点之一. 该问题旨在学习双向映射以匹配两个域的联合分布，目前仍然面临两个重要挑战：第一，两个不同域之间的相关性信息难以被充分利用. 第二，联合分布匹配问题难以建模和优化. 基于最优传输理论，本文通过最小化两个域间联合分布的 Wasserstein 距离来解决上述挑战. 首先，本文提出一个定理将难以求解的 Wasserstein 距离原问题转化为一个简单的优化问题，并设计了一个联合 Wasserstein 自编码器模型（JWAE）来求解该问题. 然后，本文将 JWAE 成功应用在无监督图像翻译和跨域视频合成任务中，并生成高质量的图像和连贯的视频. 实验结果表明，JWAE 在两种任务中的定性和定量指标上均优于现有方法. 比如，在“街景→语义分割”图像翻译任务中，JWAE 的 IS 值比 CycleGAN 高 0.59，以及 FID 值比 CycleGAN 小 65.8. 在“冬季→夏季”跨域视频合成任务中，JWAE 的 FID4video 值比 Slomo-Cycle 小 2.2.

关键词 联合分布匹配；最优传输理论；Wasserstein 距离；无监督图像翻译；跨域视频合成

中图法分类号 TP311

Joint Distribution Matching Method and Applications based on Optimal Transport Theory

CAO Jie-Zhang[1)] MO Lang-Yuan[1)] DU Qing[1)] GUO Yong[1)] ZHAO Pei-Lin[2)]

HUANG Jun-Zhou[2)] TAN Ming-Kui[1)]

[1)](School of Software Engineering, South China University of Technology, Guangzhou 510006)

[2)](Tencent AI Lab, Shenzhen 518054)

Abstract Joint distribution matching problem is one of the research hotspots in the field of machine learning and computer vision. This problem, which aims to learn bidirectional mappings to match joint distributions of two different domains, has two critical challenges: First, it is very difficult to exploit sufficient correlation information from the joint distributions of two different domains. In the unsupervised learning setting, there are two sets of samples drawn separately from two marginal distributions in two different domains. Based on the coupling theory, there exist an infinite set of joint distributions given two marginal distributions, and thus infinite bidirectional mappings between two different domains may exist. Therefore, directly learning the joint distributions without additional information between the marginal distributions is a highly ill-posed problem; Second, the joint distribution matching problem is hard to formulate and effectively optimize. One can directly apply some statistics divergence (*e.g.*, Wasserstein distance) to measure the divergence of joint distributions.

本课题得到广东省重点领域研发计划项目(2018B010107001)，国家自然科学基金重点项目(61836003)，广东省珠江人才计划创新创业团队(2017ZT07X183)，中央高校基本科研业务费专项资金(D2191240)，腾讯人工智能实验室犀牛鸟重点研究项目(JR201902)资助. 曹杰彰，硕士研究生，研究领域为对抗机器学习。邮箱：secaojiezhang@mail.scut.edu.cn. 莫朗元(**共同第一作**)，硕士研究生，研究领域为机器视觉，邮箱：selymo@mail.scut.edu.cn. 杜卿(**共同第一作**)，博士，副教授，研究领域为深度学习，邮箱：duqing@scut.edu.cn. 国雍，博士研究生，研究领域为深度学习，邮箱: guoyongcs@gmail.com. 赵沛霖，博士，研究领域为机器学习和数据挖掘，邮箱: peilinzhao@hotmail.com. 黄俊州，博士，副教授，研究领域为机器学习和机器视觉，邮箱：jzhuang@uta.edu. 谭明奎(**通信作者**)，博士，教授，研究领域为机器学习. 邮箱: mingkuitan@scut.edu.cn.

Wasserstein distance is a measure in the optimal transport theory, which has been successfully applied in computer vision applications. However, directly optimizing the primal problem of Wasserstein distance may result in intractable computational cost and statistical difficulties. Recently, many studies have been proposed to address the joint distribution matching problem and learn the mappings in two domains separately, which cannot learn cross-domain correlations and may incur joint distribution mismatching problem. In this paper, relying on the optimal transport theory, we tackled these issues by minimizing Wasserstein distance of the joint distributions in two different domains. However, directly optimizing the primal problem of Wasserstein distance is intractable due to the computational cost and the statistical difficulties. Without loss of generality, two different domains can be assumed that they often share the same latent space (*i.e.*, images in different domains have similar characteristics), we then proposed a theorem to reduce the intractable optimization problem into a simple and feasible problem. With the help of the proposed theorem, we introduced a novel objective function and design a Joint Wasserstein Auto-Encoders (JWAE) to solve the joint distribution matching problem. Our novel objective function is composed of two parts, *i.e.*, reconstruction loss and distribution divergence. The reconstruction loss can be derived from Auto-Encoder and cycle mapping, while the distribution divergence needs to be optimized for three spaces (*i.e.*, source space, target space and latent space). In this way, we can learn good bidirectional mappings through minimizing the reconstruction loss and reducing the distribution divergence. In the experiments, we applied our proposed method to perform unsupervised image-to-image translation and cross-domain video-to-video synthesis, and generate high quality images and coherent videos. Both qualitative and quantitative comparisons demonstrate the superior performance of our method over several baseline methods. For example, on the "scene→segmentation" image-to-image translation task, the IS value of JWAE is 0.59 higher than that of CycleGAN, and the FID value of JWAE is 65.8 lower than that of CycleGAN. On the "winter→summer" video synthesis, the FID4video value of JWAE is 2.2 lower than that of Slomo-Cycle.

Key words Joint distribution matching；Optimal transport theory；Wasserstein distance；Unsupervised image translation；Cross-domain video synthesis

1 引言

联合分布匹配问题旨在学习两个域之间的双向映射以匹配数据的联合分布. 该问题已经成功应用在机器学习和计算机视觉中，如图像翻译[1,2]和视频合成[3,4]. 但是，解决该问题面临两个重要挑战.

第一个重要挑战是如何充分利用两个域的联合分布信息. 在无监督学习中，两个域的数据分别取于各自的边缘分布. 根据概率耦合理论[5]，给定两个边缘分布，存在无限多个联合分布的集合，即在两个不同域之间可能存在无限多个双向映射. 因此，当边缘分布之间不存在额外信息时，直接学习联合分布是一个病态问题. 最近，许多研究[1,6,7]已经被提出来分别学习两个域中的映射，但是它们无法学习跨域相关性信息. 因此，如何从联合分布中获取足够的信息仍然是一个开放性问题.

第二个重要挑战是如何去构造和优化联合分布匹配问题. 现有方法[8,9]无法直接度量联合分布之间的距离，这可能会导致联合分布不匹配问题. 一种简单的方法是直接利用一些统计度量距离（如 Wasserstein 距离）来解决这个问题. 但是，直接优化这些度量距离可能会带来不可接受的计算代价和统计困难[10]. 所以，如何为联合分布匹配问题设计新的目标函数和有效的优化方法非常重要.

针对上述两个重要挑战，本文提出了一个联合 Wasserstein 自编码器（JWAE）模型，通过最小化联合分布之间的 Wasserstein 距离来学习两个域之间的双向映射. 具体而言，为了解决第一个挑战，本文使用最优传输理论来利用两个不同域之间的几何信息和相关性. JWAE 能够最小化两个联合分布的距离，从而可以利用两个域的联合分布信息. 对于第二个挑战，JWAE 利用最优传输理论对联合分布匹配问题进行建模，即最小化不同域的两个联合分布之间的 Wasserstein 距离. 然后，本文从理论上分析和推导出一个易于优化的定理，并提出一个新的目标函数进行优化，从而避免求解原问题所带来的严重计算代价和优化困难.

本文主要贡献总结如下：

- 基于最优传输理论，本文提出 JWAE 来解决联

合分布匹配问题. 该方法能够学习和获取不同域之间的相关性信息，

- 在理论上，本文推导出一个重要的定理，使得最小化联合分布之间 Wasserstein 距离这个难以解决的问题转化为一个简单的优化问题.
- 在实验上，本文将 JWAE 应用于无监督图像翻译和跨域视频合成任务上. 实验表明 JWAE 在两个任务中均优于当前代表性的对比方法.

2 相关工作

（1）图像翻译

生成对抗网络(GAN)[11-13]、变分自编码器[14]和 Wasserstein 自编码器[15]成功应用在深度学习中[16]，如数据生成和图像翻译[17]任务. 例如，Pix2pix[18]利用条件 GAN 来翻译图像，但是难以生成高分辨率的图像. Wang 等人[19]设计一个新的对抗损失函数来生成高分辨率的图像. CycleGAN[1]和 DualGAN[7]通过最小化两个域中的对抗损失和循环一致性损失来学习跨域映射. SCAN[20]提出堆叠式的 GAN 来翻译高分辨率的图像. Gokaslan 等人[21]通过设计判别器来训练生成器. 最新的研究，如 HarmonicGAN[22]，引入了一个光滑项和学习一致性的映射，以学习更好的跨域映射. 此外，Alami 等人[23]将无监督注意力机制引入到 GAN 中来提高图像翻译的质量. 但是，它们常常忽略去解决联合分布匹配问题，从而无法学习域间的相关性信息，导致图像细节或轮廓的清晰度不够.

（2）跨域视频合成

联合分布匹配可应用于跨域视频合成，所以本文进一步研究跨域视频合成问题. 由于大部分图像翻译方法[1,6,7]无法在视频中进行插值而不能直接合成跨域视频，为此，本文结合了一些视频帧插值方法[24-27]来合成视频. 此外，UNIT[2]在潜在空间中进行插值来合成视频，但是合成的视频质量不高. 最近，视频翻译方法[3]将视频从一个域翻译到另一个域，但是它不能进行视频帧插值，因此不能应用在跨域视频合成中. 此外，Vid2vid[4]是一个视频到视频合成方法，但是不能直接应用在无监督学习的任务中. 因此，这些方法无法匹配不同域的联合分布，导致合成视频存在质量不好和连续性差等问题.

（3）联合分布匹配.

最近，CoGAN[28]利用参数共享约束来学习联合分布. 由于该方法以噪声为输入，从而无法控制输出. UNIT[2]假设共享的潜在空间和参数共享来学习联合分布. BiGAN[29]学习一个映射及其反映射来学习隐变量和数据的联合分布. ALI[30]学习生成网络和推理网络使得编码器联合分布和解码器联合分布相互逼近. ALICE[8]通过研究联合分布匹配问题来理解和提升对抗学习的性能. 这些方法通过学习和匹配联合分布，目标是在生成任务上提升 GAN 的性能. 然而，它们难以直接拓展到两个域的图像翻译和视频合成任务中的联合分布匹配.

3 预备知识

本文使用花体字母 X 来定义空间，大写字母 X 来定义随机变量，以及黑体小写字母 $\mathbf{x}$ 来定义随机变量相应的值. 令 $P(X)$ 为概率分布，其密度函数用 $p(\mathbf{x})$ 表示.令 (X, P_X) 为一个域，P_X 为 X 中的边缘分布，以及 $\mathrm{P}(\mathrm{X})$ 为 X 中所有概率测度的集合.

生成对抗网络(GAN)[11]通过极小极大化游戏来训练生成模型. 它包含一个生成器 G 和一个判别器 D. 其中，生成器 G 为了学习真实数据分布，而判别器 D 为了判别输入数据是来自真实数据还是生成器. 具体而言，生成对抗网络的目标函数定义为：

$$\min_G \max_D \mathop{\mathrm{E}}_{\mathbf{x}\in P_X}\left[\log D(\mathbf{x})\right] + \mathop{\mathrm{E}}_{\mathbf{z}\in P_Z}\left[\log(1-D(G(\mathbf{z})))\right],$$

其中，P_X 是一个真实数据分布，P_Z 是一个先验分布，如高斯分布和均匀分布.

基于最优传输理论[31]，给定一个真实分布 P_X 和生成分布 P_G，通过学习一个联合分布 $P\in\mathrm{P}(P_X, P_G)$，Wasserstein 距离定义为

$$\mathrm{W}(P_X, P_G) = \inf_{P\in\mathrm{P}(X,X')} \mathrm{E}_P\left[c(X, X')\right], \qquad (1)$$

其中 $c(X, X')$ 是一个距离函数，$\mathrm{P}(X\sim P_X, X'\sim P_G)$ 是关于边缘分布 P_X 和 P_G 所有联合分布的集合. 在实验中，直接优化原问题(1)难度大且需要极大的优化代价. 为此，WGAN[32]通过优化对偶问题来提升生成任务的性能. 最近，Lei 等人[33]从几何的角度对 WGAN 等方法进行了分析，但是这些分析难以帮助求解联合分布匹配问题. 为此，本文从理论上提出一个数学定理将难以求解的原问题转化为一个简单的优化问题.

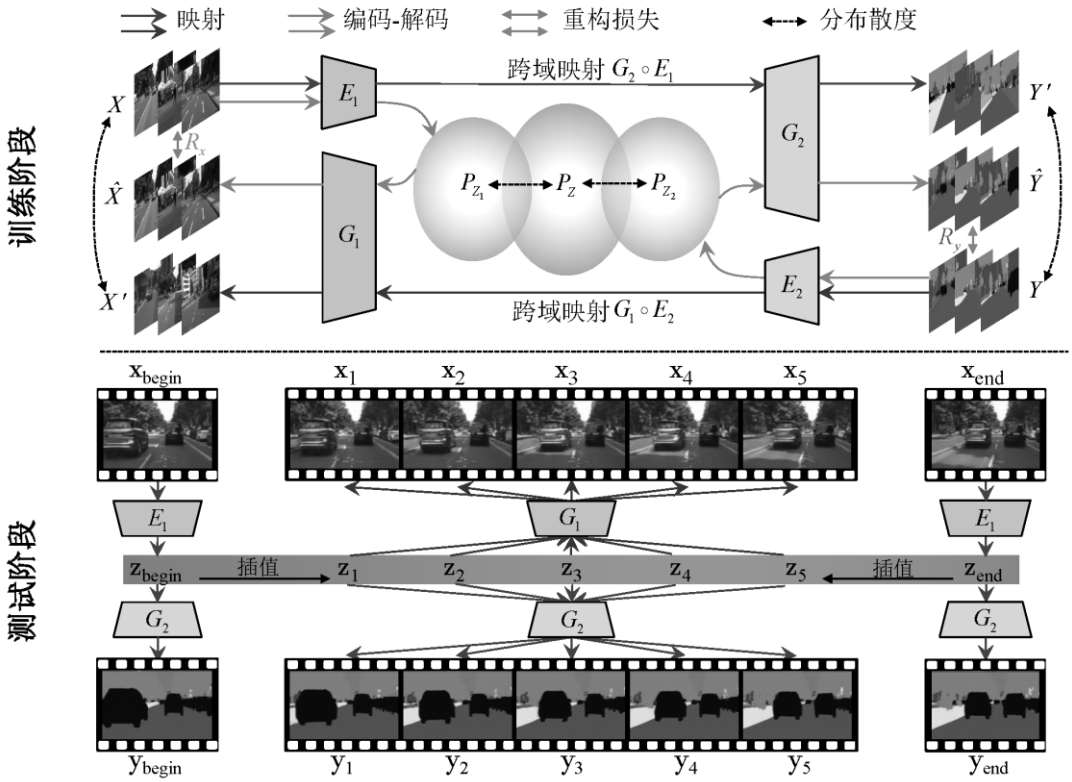


图 1 联合 Wasserstein 自编码器在无监督图像翻译和跨域视频合成任务上的训练和测试阶段. 训练阶段：给定真实数据 X 和 Y，本文学习跨域映射（$G_2 \circ E_1$ 和 $G_1 \circ E_2$）来生成样本 Y' 和 X'，使得生成分布最终能够拟合真实分布. 另外，通过自编码器 $G_1 \circ E_1$ 和 $G_2 \circ E_2$ 编码产生的潜在空间分布也应该相互逼近. 测试阶段：给定来自源域的两个视频帧 $\mathbf{x}_{begin}$ 和 $\mathbf{x}_{end}$，本文提取其潜在变量 $\mathbf{z}_{begin}$ 和 $\mathbf{z}_{end}$，然后通过线性插值和跨域映射来合成两个域的相应视频. 更多细节可以关注章节 4.3.

4 联合 Wasserstein 自编码器

4.1 问题定义

本文旨在解决联合分布匹配问题，该问题已经在机器学习和计算机视觉任务中得到广泛应用，例如无监督图像到图像翻译[1,2]和无监督跨域视频合成[3,4]. 然而，该问题难以建模和进一步优化求解. 因此，本文首先对联合分布和联合分布匹配问题给出一个正式的数学定义.

在无监督学习中，两个域的数据分别取自这两个域的边缘分布. 具体而言，令 (X, P_X) 和 (Y, P_Y) 表示两个域，其中 P_X 和 P_Y 分别是在 X 和 Y 上的边缘分布. 本文目的是为了学习两个跨域映射 $f:\mathrm{X}\to\mathrm{Y}$ 和 $g:\mathrm{Y}\to\mathrm{X}$，因此定义如下联合分布：

$$P_{\mathrm{A}}(X,Y') \quad 和 \quad P_{\mathrm{B}}(X',Y), \tag{2}$$

其中 $Y'=f(X)$ 和 $X'=g(Y)$. 本文目的是匹配联合分布 P_{A} 和 P_{B}，即 P_{A} 和 P_{B} 相互逼近.

本文给出联合分布匹配的问题定义. 首先基于最优传输理论中的 Wasserstein 距离（公式(1)），本文度量两个联合分布 $P_{\mathrm{A}}(X,Y')$ 和 $P_{\mathrm{B}}(X',Y)$ 之间的分布距离，即 $\mathrm{W}_c(P_{\mathrm{A}},P_{\mathrm{B}})$. 给定两个域 (X,P_X) 和 (Y,P_Y)，数据 $Y'=f(X)$ 由映射 f 生成，而 $X'=g(Y)$ 由映射 g 生成. 本文的目的是让 $P_{\mathrm{A}}(X,Y')$ 和 $P_{\mathrm{B}}(X',Y)$ 相互逼近，即让它们之间的分布距离最小，然后学习两个跨域映射 f 和 g 来匹配该两个联合分布 $P_{\mathrm{A}}(X,Y')$ 和 $P_{\mathrm{B}}(X',Y)$. 具体而言，为了匹配这两个联合分布，本文通过最小化 Wasserstein 距离来学习两个跨域映射 f 和 g，即：

$$\mathrm{W}_c(P_{\mathrm{A}},P_{\mathrm{B}})=\min_{P\in\mathrm{P}(P_{\mathrm{A}},P_{\mathrm{B}})}\mathrm{E}_P\left[c(X,Y',X',Y)\right], \tag{3}$$

其中联合分布集 $\mathrm{P}(P_{\mathrm{A}},P_{\mathrm{B}})$ 是由所有联合分布组成的集合，其中 P_{A} 和 P_{B} 皆为边缘分布. 这里，$c(X,Y';X',Y)$ 是距离函数，用于度量 (X,Y') 和 (X',Y) 的距离，如欧几里得距离.

4.2 联合Wasserstein分布匹配

在优化中，直接求解问题(3)会面临两个挑战. 第一，直接优化该问题会产生严重的计算代价[10]. 第二，如何选择一个合适的距离函数非常困难. 为了克服上述挑战，本文通过下面的定理将问题(3)转化为一个更简单的优化问题.

定理 1. 给定解码器 G_1 和 G_2，并假设其确定性模型 $P_{G_1}(X'|Z)$ 和 $P_{G_2}(Y'|Z)$ 是狄拉克函数，定义为 $P_{G_1}(X'|Z=\mathbf{z})=\delta_{G_1(\mathbf{z})}$ 和 $P_{G_2}(Y'|Z=\mathbf{z})=\delta_{G_2(\mathbf{z})}$，对于 $\forall \mathbf{z}\in \mathrm{Z}$，则 P_{A} 和 P_{B} 的 Wasserstein 距离可以定义为:

$$\begin{aligned}\mathrm{W}_c\left(P_{\mathrm{A}},P_{\mathrm{B}}\right)=&\inf_{Q\in\mathrm{Q}_1}\mathrm{E}_{P_X}\mathrm{E}_{Q(Z_1|X)}\left[c_1\left(X,G_1\left(Z_1\right)\right)\right]\\&+\inf_{Q\in\mathrm{Q}_2}\mathrm{E}_{P_Y}\mathrm{E}_{Q(Z_2|Y)}\left[c_2\left(G_2\left(Z_2\right),Y\right)\right],\end{aligned}\tag{4}$$

其中定义两个集合 $\mathrm{Q}_1=\{Q(Z_1|X)|Q\in\tilde{\mathrm{Q}},P_Y=Q_Y\}$ 和 $\mathrm{Q}_2=\{Q(Z_2|Y)|Q\in\tilde{\mathrm{Q}},P_X=Q_X\}$ 为所有编码器的集合，而 Q 属于 $\tilde{\mathrm{Q}}=\{Q|P_{Z_1}=Q_{Z_1},P_{Z_2}=Q_{Z_2}\}$.

证明请参阅附录. 注意定理 1 中使用狄拉克函数，因为现实中数据分布是离散的，狄拉克函数常用于对这些离散分布来表示和建模[10,34].

由定理 1 可知，我们的目标是优化两个自编码器的重构损失函数使得生成分布接近真实分布. 具体而言，令 $\mathrm{R}_x(F)$ 和 $\mathrm{R}_y(F)$ 是两个关于自编码器（即 $G_1\circ E_1$ 和 $G_2\circ E_2$）的重构损失函数，和令 $F=\{E_1,E_2,G_1,G_2\}$，本文优化问题(4)为：

$$\begin{aligned}&\min_F \mathrm{W}_c\left(P_{\mathrm{A}},P_{\mathrm{B}}\right):=\mathrm{R}_x(F)+\mathrm{R}_y(F)\\&s.t.\ P_X=Q_X,\ P_Y=Q_Y,\ P_Z=Q_Z,\end{aligned}\tag{5}$$

其中 P_X，P_Y 和 P_Z 是真实分布，Q_X，Q_Y 和 Q_Z 是生成分布. 通过优化问题(5)，生成分布将接近真实分布.

基于定理 1，本文将问题(3)转变为了一个简单的优化问题. 为了优化问题(5)，本文加入对生成和真实分布的约束项（$P_X=Q_X$，$P_Y=Q_Y$ 和 $P_Z=Q_Z$）. 给定自编码器（$G_1\circ E_1$ 和 $G_2\circ E_2$）和模型 $F=\{E_1,E_2,G_1,G_2\}$，我们优化以下问题：

$$\begin{aligned}\min_F \mathrm{W}_c\left(P_{\mathrm{A}},P_{\mathrm{B}}\right):=&\mathrm{R}_x(F)+\mathrm{R}_y(F)+\lambda_x d\left(P_X,Q_X\right)\\&+\lambda_y d\left(P_Y,Q_Y\right)+\lambda_z d\left(P_Z,Q_Z\right),\end{aligned}\tag{6}$$

其中 $d(\cdot,\cdot)$ 是分布距离，而 λ_x，λ_y 和 λ_z 是超参数. 注意，本文的目标函数包含两种类型的损失函数，即重构损失（即 $\mathrm{R}_x(F)$ 和 $\mathrm{R}_y(F)$）和分布散度（即 $d(P_X,Q_X),d(P_Y,Q_Y)$ 和 $d(P_Z,Q_Z)$）. 下面本文将详细定义这两种损失函数.

（1）重构损失.

为了优化目标函数(6)，重构损失必须要很小，这意味着任何输入的重构输出必须足够接近于源域和目标域的输入. 如图 1 所示，来自源域的 $\mathbf{x}$ 的重构输出可以由自编码器重构 $G_1(E_1(\mathbf{x}))$ 和循环映射 $G_1(E_2(G_2(E_1(\mathbf{x}))))$ 两部分组成. 以源域为例，给定一个输入 $\mathbf{x}$，本文优化以下重构损失：

$$\begin{aligned}\mathrm{R}_x(F)=\hat{\mathrm{E}}_{\mathbf{x}\in P_X}\Big[&\left\|\mathbf{x}-G_1(E_1(\mathbf{x}))\right\|_1\\&+\left\|\mathbf{x}-G_1(E_2(G_2(E_1(\mathbf{x}))))\right\|_1\Big],\end{aligned}\tag{7}$$

其中，$\hat{\mathrm{E}}$[是经验期望. 注意，公式(7)第一项是自编码器重构损失函数，第二项是循环一致损失函数[1]. 对于目标域而言，$\mathrm{R}_y(F)$ 以同样的方式构造.

（2）分布散度.

从定理 1 得，其约束强制生成的分布应该等于源域和目标域中的真实分布. 而且，由两个编码器（即 $G_1\circ E_1$ 和 $G_2\circ E_2$）生成的潜在分布应该接近于一个先验的潜在分布. 因此本文需优化三个分布散度（即 $d(P_X,Q_X),d(P_Y,Q_Y)$ 和 $d(P_Z,Q_Z)$），如图 1 所示. 值得注意的是，本文不对分布散度的选择作任何限制，例如对抗损失（原始 GAN 或 WGAN）或者最大平均差异（MMD）. 在实验中，本文使用原始 GAN 来度量真实和生成分布的距离. 以 $d(P_X,Q_X)$ 为例，本文优化以下的损失函数：

$$\begin{aligned}d\left(P_X,Q_X\right)=\max_{D_x}\Big[&\hat{\mathrm{E}}_{\mathbf{x}\sim P_X}\left[\log D_x\left(\mathbf{x}\right)\right]\\&+\hat{\mathrm{E}}_{\tilde{\mathbf{x}}\sim Q_X}\left[\log\left(1-D_x\left(\tilde{\mathbf{x}}\right)\right)\right]\Big],\end{aligned}\tag{8}$$

其中 $\mathbf{x}$ 表示从真实分布 P_X 采样的样本，$\tilde{\mathbf{x}}$ 表示从生成分布 Q_X 采样的样本，而 D_x 是一个关于源域的判别器. 同样地，另外两个损失函数 $d(P_Y,Q_Y)$ 和 $d(P_Z,Q_Z)$ 可理可得，参见补充材料（见附件 1）.

4.3 联合Wasserstein自编码器的应用

联合 Wasserstein 自编码器可以应用在无监督图像翻译与无监督跨域视频合成任务上.

（1）无监督图像翻译.

如图 1 所示，JWAE 能够解决无监督图像翻译问题. 具体而言，给定来自源域和目标域的真实数据 $\mathbf{x}$ 和 $\mathbf{y}$，联合 Wasserstein 自编码器学习跨域映射 $G_2\circ E_1$ 和 $G_1\circ E_2$ 来生成样本 $\mathbf{y}'$ 和 $\mathbf{x}'$. 通过最小化目标函数 $\mathrm{W}_c(P_{\mathrm{A}},P_{\mathrm{B}})$，可以学习如下跨域映射：

$$\mathbf{y}'=G_2\circ E_1(\mathbf{x})\quad 和\quad \mathbf{x}'=G_1\circ E_2(\mathbf{y}),\tag{9}$$

其中 E_1 和 E_2 是编码器，G_1 和 G_2 是解码器. 具体的训练细节如算法 1 所示.

算法 1. JWAE 训练算法.

输入：两个域训练数据 $\{\mathbf{x}_i\}_{i=1}^{M}$ 和 $\{\mathbf{y}_k\}_{k=1}^{N}$，初始化所有模型

输出：编码器 E_1, E_2 和解码器 G_1, G_2

REPEAT

通过梯度上升更新 D_x, D_y, D_z：

$$\lambda_x d(P_X, Q_X) + \lambda_y d(P_Y, Q_Y) + \lambda_z d(P_Z, Q_Z)$$

通过梯度下降更新 E_1, E_2, G_1, G_2：

$$\mathrm{R}_x(F) + \mathrm{R}_y(F) + \lambda_x d(P_X, Q_X) + \lambda_y d(P_Y, Q_Y) + \lambda_z d(P_Z, Q_Z)$$

UNTIL 模型收敛

（2）跨域视频合成.

JWAE 可用在跨域视频合成任务. 给定来自源域的两个输入视频帧 $\mathbf{x}_{begin}$ 和 $\mathbf{x}_{end}$，本文在从 $\mathbf{x}_{begin}$ 和 $\mathbf{x}_{end}$ 编码得到的潜在变量之间进行线性插值. 接着插值得到的潜在变量被解码成源域和目标域的相应视频帧（见图 1 和算法 2）. 两个域的插值如下：

$$\mathbf{x}_{mid} = G_1(\mathbf{z}_{mid}) \quad 和 \quad \mathbf{y}_{mid} = G_2(\mathbf{z}_{mid}), \tag{10}$$

其中 $\mathbf{z}_{mid} = \rho E_1(\mathbf{x}_{begin}) + (1-\rho) E_1(\mathbf{x}_{end}), \rho \in (0,1)$，$\mathbf{x}_{mid}$ 表示在源域的插值帧，以及 $\mathbf{y}_{mid}$ 表示在目标域的插值帧. 具体过程由算法 2 所示.

算法 2. 跨域视频合成算法.

输入：源域的测试数据: $\mathbf{x}_{begin}$ 和 $\mathbf{x}_{end}$，插值帧数 n

输出：合成的源域和目标域视频 $\{\mathbf{x}_{begin}, \{\mathbf{x}_{mid}\}_{mid=1}^{n}, \mathbf{x}_{end}\}$ 和 $\{\mathbf{y}_{begin}, \{\mathbf{y}_{mid}\}_{mid=1}^{n}, \mathbf{y}_{end}\}$

步骤 1：线性插值合成源域视频序列

$$\mathbf{z}_{begin} = E_1(\mathbf{x}_{begin}), \mathbf{z}_{end} = E_1(\mathbf{x}_{end})$$

$$\mathbf{z}_{mid} = \rho \mathbf{z}_{begin} + (1-\rho)\mathbf{z}_{end}, \rho \in \{1/n, \ldots, (n-1)/n\}$$

$$\mathbf{x}_{mid} = G_1(\mathbf{z}_{mid})$$

合成的视频：$\{\mathbf{x}_{begin}, \{\mathbf{x}_{mid}\}_{mid=1}^{n}, \mathbf{x}_{end}\}$

步骤 2：翻译合成目标域视频序列

$$\mathbf{y}_{begin} = G_2(\mathbf{z}_{begin}), \mathbf{y}_{mid} = G_2(\mathbf{z}_{mid}), \mathbf{y}_{end} = G_2(\mathbf{z}_{end})$$

合成的视频：$\{\mathbf{y}_{begin}, \{\mathbf{y}_{mid}\}_{mid=1}^{n}, \mathbf{y}_{end}\}$

4.4 联合Wasserstein自编码器和WGAN的区别

- 研究问题不同：JWAE 解决联合分布匹配问题，而 WGAN 解决生成和真实数据分布匹配问题.
- 动机与目的不同：JWAE 旨在学习如何利用域间的相关性信息和优化联合分布匹配问题. 然而，直接利用 WGAN 难以解决这些挑战.
- 目标函数不同：两种方法都是基于最优传输理论，JWAE 的目标函数包含自编码损失函数和分布距离函数，而 WGAN 只有分布距离.

5 实验

本文实现细节如下：JWAE 的自编码器网络结构（$G_1 \circ E_1$ 和 $G_2 \circ E_2$）由 CycleGAN[1]的生成器网络结构修改而来，首先是两个步长为 2 的卷积层进行下采样，中间有六个残差模块，后接两个步长为 2 的卷积层进行上采样. 对于判别器，本文采用 PatchGAN[18]的网络结构，有关网络结构的更多详细信息，请参阅补充材料（见附件 1）. 本文遵循 CycleGAN 的实验设定，并选用 Adam 优化器[35]. 训练过程中将图片输入尺寸统一调整到 128×128 的分辨率，小批次数量为 8，学习率为 0.0002，训练轮数为 200 轮. 前 100 轮学习率不变，后 100 轮学习率线性衰减到 0. 对于公式(6)中的超参数 λ_x 和 λ_y，参照 CycleGAN[1]，本文设置超参数 $\lambda_x = \lambda_y = 0.1$ 和 $\lambda_z = 0.1$（超参数 λ_z 的敏感度分析见章节 5.4）.

本文使用以下两个数据集：

- Cityscapes[36]包含 2900 张画面连续的德国城市街景图片及其相对应的语义分割. 训练集与测试集划分为 2600 张和 300 张.
- SYNTHIA[37]包含许多不同场景、不同季节（即春夏秋冬）的卡通合成视频. 本实验选取一个场景四季的视频，然后进行“冬季→{春季，夏季，秋季}”图像翻译. 单个季节的训练集和测试集数量分别为 2140 张和 240 张.

本文使用以下评价指标：

- Inception Score (IS)[38]广泛用于生成模型中. IS 通过利用 Inception-V3 模型[39]的类别预测信息来评估生成样本的质量和多样性.
- Fréchet Inception Distance (FID)[40]也是一个广泛使用在生成模型上的指标. FID 可以评估生成图像的质量，因为它能捕获生成样本与真实样本的相似性，并与人类判断相关联.
- Video variant of FID (FID4Video) [41]评估视频的质量和连贯性. 本文使用一个预训练的视频识别模型 I3D[41]对视频序列提取特征. 然后，对这些特征计算 FID4Video.

一般而言，对于 IS 指标，值越大代表着翻译的图像质量越好；对于 FID 和 FID4Video 这两种指标，值越小意味着翻译的图像或视频的质量越好.

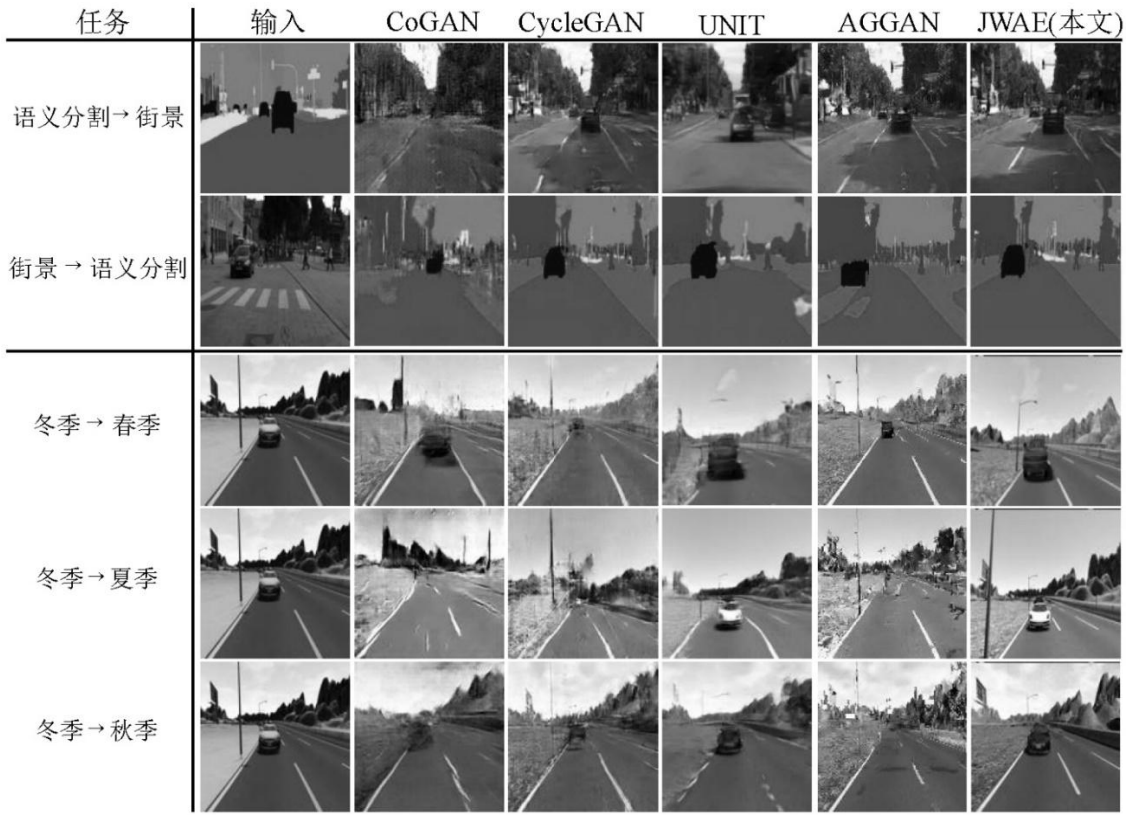


图 2　无监督图像到图像翻译任务：不同方法在 Cityscapes 和 SYNTHIA 数据集上的翻译效果比较.

表 1　无监督图像到图像翻译任务：不同方法在 Cityscapes 和 SYNTHIA 数据集上的指标比较.

方法	街景→语义分割		语义分割→街景		冬季→春季		冬季→夏季		冬季→秋季	
	IS	FID	IS	FID	IS	FID	IS	FID	IS	FID
CoGAN[28]	1.76	230.47	1.41	334.61	2.13	314.63	2.05	372.82	2.28	300.47
CycleGAN[1]	1.83	87.69	1.70	124.49	2.23	115.43	2.32	120.21	2.30	100.30
UNIT[2]	2.01	65.89	1.66	89.79	2.55	88.26	2.41	89.92	2.46	85.26
AGGAN[23]	1.90	126.27	1.66	115.87	2.12	140.97	2.02	152.02	2.32	124.38
JWAE (本文)	**2.42**	**21.89**	**1.92**	**42.13**	**3.12**	**82.34**	**2.86**	**84.37**	**2.85**	**83.26**

5.1 无监督图像翻译任务中的结果

本文进行无监督图像到图像翻译任务，并将 JWAE 与以下对比方法进行比较.

- CoGAN[28]通过强制权重共享的约束，然后学习两个不同域的联合分布.
- CycleGAN[1]结合对抗损失与循环一致损失，进而提升图像翻译的性能.
- UNIT[2]使用了一个共享潜在空间的假设来学习不同域中图像的联合分布.
- AGGAN[23]引入了一个无监督的注意力机制. AGGAN 将一个注意力网络引入到生成器中来翻译出更加准确的图像.

本文在 Cityscapes 和 SYNTHIA 上做五个图像翻译实验，分别是“街景→语义分割”、“语义分割→街景”、“冬季→{春季，夏季，秋季}”. 定量和定性结果比较分别可见表 1 和图 2.

对于定量比较（见表 1），JWAE 在两种指标上均最优，即 IS 值最高和 FID 值最低. 说明生成图像的质量最高，并证明它能够有效地学习到域间的联合分布. 对于其他方法，CoGAN 的指标最差，因为它无法直接对输入图像进行翻译，而是采样潜在变量来生成目标域的图像. AGGAN 和 CycleGAN 的指标稍差，因为它们不能学习到跨域相关性信息，因此翻译的图像可能会丢失部分信息. 由于 UNIT 难以学习到很好的联合分布，所以效果比 JWAE 差.

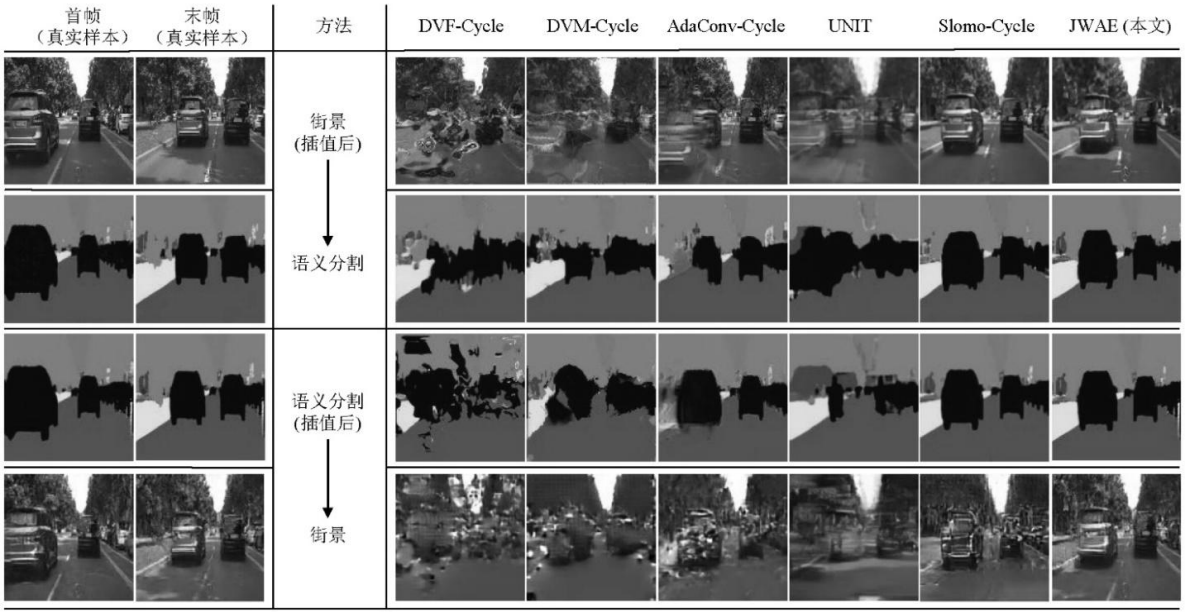


图 3 不同方法在 Cityscapes 数据集上的跨域视频合成结果比较. 本实验首先合成街景的视频，然后将其翻译到语义分割域（上半部分）. 同样，本实验也进行语义分割到街景的翻译（下半部分）.

表 2 无监督跨域视频合成：不同方法在 Cityscapes 和 SYNTHIA 数据集上的指标比较.

方法	街景→语义分割			语义→分割街景			冬季→春季			冬季→夏季			冬季→秋季		
	IS	FID	FID4Video	IS	FID	FID4Video	IS	FID	FID4Video	IS	FID	FID4Video	IS	FID	FID4Video
DVF-Cycle	1.43	110.59	23.95	1.34	151.27	40.61	2.09	152.44	42.22	2.07	160.69	42.43	2.44	163.13	41.04
DVM-Cycle	1.36	50.51	17.33	1.26	116.62	40.83	1.98	129.80	38.19	1.99	140.86	36.66	2.19	129.02	36.64
AdaCon-Cycle	1.29	33.50	14.96	1.27	99.67	30.24	1.91	117.40	23.83	2.10	126.01	20.62	2.18	110.52	16.77
UNIT	1.66	31.27	10.12	1.89	76.72	29.21	2.14	96.40	23.12	2.13	108.01	24.70	2.30	97.73	20.39
Slomo-Cycle	**1.84**	27.35	8.71	1.89	59.21	27.87	2.27	93.77	21.53	2.41	96.27	20.19	2.36	94.41	15.65
JWAE (本文)	1.69	**22.74**	**6.80**	**1.97**	**43.48**	**25.87**	**2.36**	**88.24**	**21.37**	**2.46**	**77.12**	**17.99**	**2.50**	**87.50**	**14.14**

定性比较（见图 2），JWAE 在生成的图像中质量最高. 对于其他方法，CoGAN 翻译的结果最差，而且翻译图像中含有很多噪音. AGGAN 和 CycleGAN 的翻译效果稍差，翻译的图像丢失部分信息. UNIT 的翻译结果好于以上三种对比方法，但是它的翻译图像清晰度不足. JWAE 翻译结果的清晰度最高，这说明 JWAE 能够学习到良好的联合分布，通过利用更多的跨域信息生成了更清晰更准确的翻译图像.

5.2 跨域视频合成任务中的结果

本文通过在两个不同域之间进行线性插值和翻译将联合 Wasserstein 自编码器应用在跨域视频合成任务中. 在实验中，一个视频序列包含连续的九帧. 本文使用视频序列的第一帧和最后一帧作为输入，然后同时在两个域中插值和翻译出序列中的七帧，从而合成一个完整的视频序列.

现有的无监督图像翻译方法（除了 UNIT[2]以外），都无法在两帧之间合成中间帧，因此它们不能直接合成跨域视频. 为此，本文选用几种视图合成方法在一个域的两帧之间合成视频帧，再使用 CycleGAN[1]将源域视频序列翻译到另一个域. 在视频合成实验中，本文考虑以下对比方法：

- UNIT[2]是一种无监督图像到图像翻译的方法. 由于 UNIT 存在一个共享潜在空间，所以可以用于视频合成. 具体而言，UNIT 可以在两个域的潜在变量之间进行插值，然后合成视频序列.
- DVF-Cycle 由视图合成方法 DVF[24] 和 CycleGAN 组成. DVF 结合光流方法与神经网络方法来合成视频帧. 本文先用 DVF 在一个域里进行视频插值，然后用 CycleGAN 将插值得到的视频翻译到另一个域. 为方便起见，本文称该方法为 DVF-Cycle.

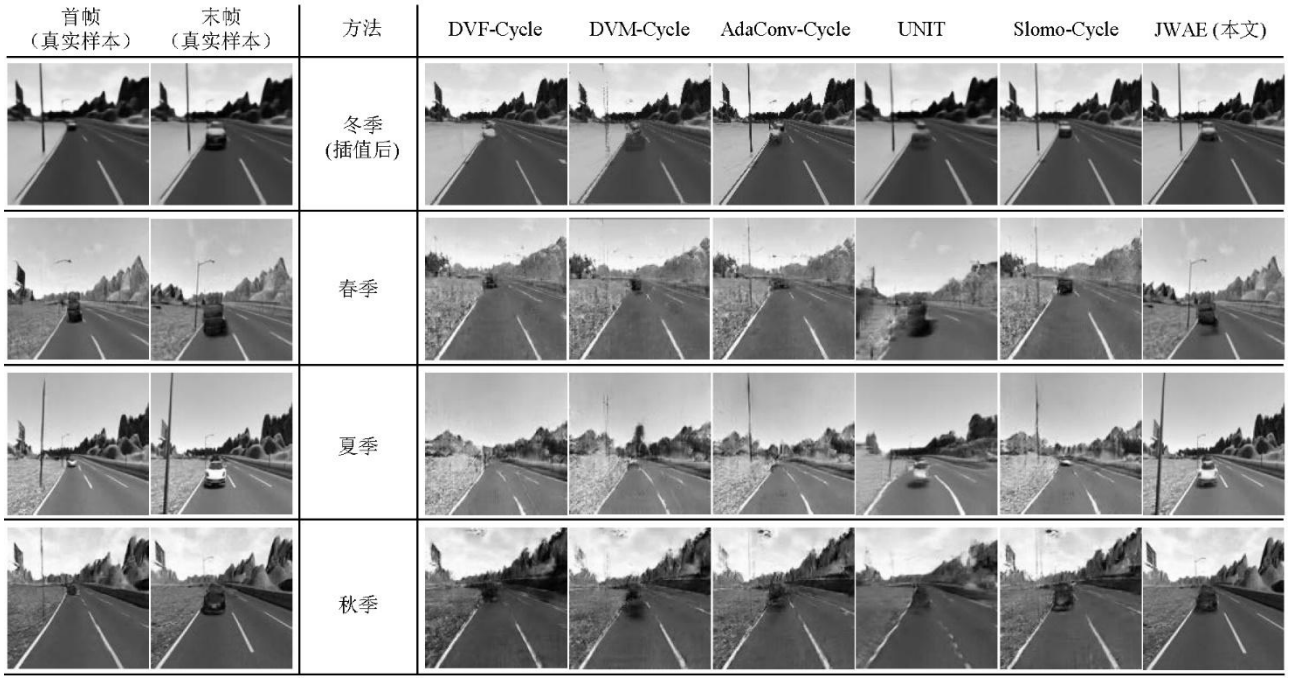


图 4　不同方法在 SYNTHIA 数据集上的跨域视频合成结果比较. 第 1 行：冬季域的合成视频. 第 2–4 行：其他三个季节（春季、夏季、秋季）相对应的翻译视频.

- DVM-Cycle 由深度视图合成方法 DVM[25]和 CycleGAN 组成. 首先，DVM 用一个校正网络校正两张输入图片，然后通过一个编码-解码网络和视图变形网络合成中间的视图. 在实验中，DVM 首先合成一个域的视频，然后 CycleGAN 将合成的视频翻译到另一个域. 为方便起见，本文称该方法为 DVM-Cycle.
- AdaConv-Cycle 由视图合成方法 AdaConv[26]和图像翻译方法 CycleGAN 组成. AdaConv 使用 1D 卷积核对将帧插值表示为输入帧上的局部可分卷积，并设计了一个深度全卷积神经网络来合成中间的视频帧. AdaConv 方法先在一个域里进行视频插值，然后 CycleGAN 将插值得到的视频翻译到另一个域. 为方便起见，本文称该方法为 AdaConv-Cycle.
- Slomo-Cycle 由视频帧合成方法 Super Slomo[27]和无监督图像翻译方法 CycleGAN 组成. 具体而言，Super Slomo 方法是一种用于可变长度多帧视频插值的端到端卷积神经网络，可以用于合成任意高帧率的视频序列. 在跨域视频合成实验中，本文先用 Super Slomo 在源域中进行多个视频帧插值，然后再利用 CycleGAN 将合成的视频翻译到目标域. 为方便起见，本文称该方法为 Slomo-Cycle.

本文的定量和定性比较情况如下：

对于定量比较，本文使用 IS、FID 与 FID4Video 三种指标评估不同方法在 Cityscapes 和 SYNTHIA 的视频合成结果，如表 2 所示. 在 IS 指标上，JWAE 在“街景→语义分割”取得不错的结果，而在另外四个任务中取得最好的结果. 说明 JWAE 插值与翻译的视频帧的质量总体上好于其他方法. 在 FID 指标上，JWAE 在所有实验上均取得最优结果. 相对于其他方法，JWAE 插值与翻译的视频帧的质量更好，更符合人类的视觉效果. 同样地，JWAE 在 FID4Video 指标上始终优于其他对比方法，说明它能够利用更多的联合分布信息生成高质量的视频帧和连贯的视频.

对于定性比较，本文对比 JWAE 和其他方法在 Cityscapes 和 SYNTHIA 上的定性结果.[1]

第一，Cityscapes 数据集上的定性分析如下：本文把街景和语义分割依次作为源域，首先在源域进行视频插值，然后将其翻译到目标域. 在图 3 中，本文比较了插值产生的视频帧与翻译得到的视频帧的定性结果. 由图 3 第一和第三行可知，DVF-Cycle 和 DVM-Cycle 插值产生的帧质量较差. 这可能是由于 DVF 不能很好捕获数据的光流信息，而 DVM 难以处理较复杂的图像.

[1] 由于页面限制，更多视觉对比结果请参阅补充材料（见附件 1）.

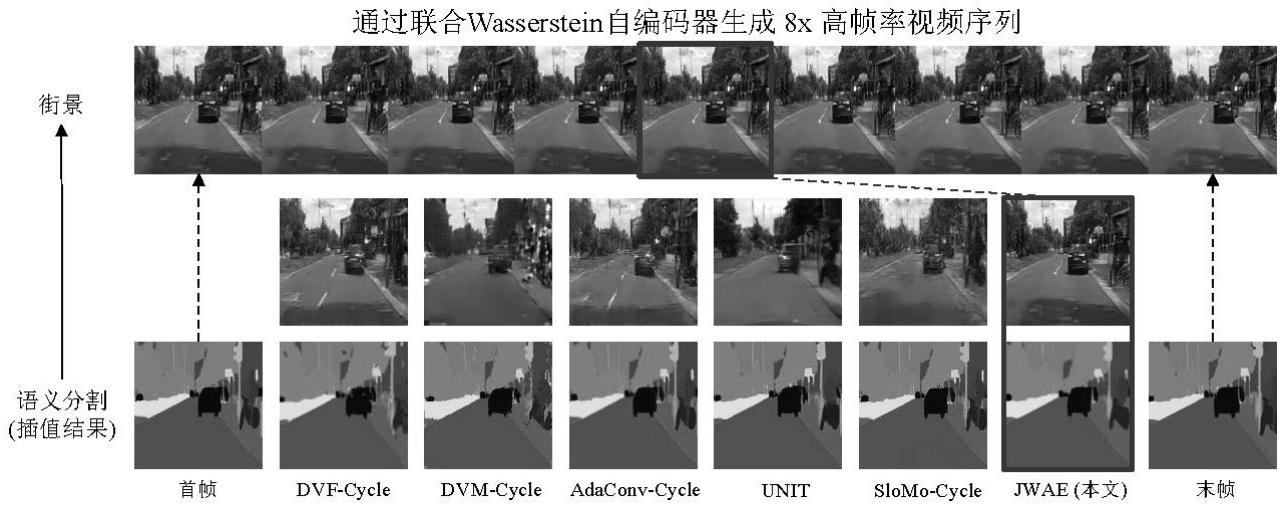


图 5 不同方法在高帧率视频合成任务上的比较，语义分割→街景. 首行：通过联合 Wasserstein 自编码器生成的 8×高帧率视频序列. 中间行：不同方法在目标域中的插值帧. 底行：不同方法在源域中的插值帧.

AdaConv-Cycle 能够合成高清的插值结果，但是无法完全消除前后帧图像变形的影响. Slomo 同时考虑了时间和空间连贯性，因而插值出质量较好的视频帧. 此外，UNIT 难以学习很好的潜在空间导致插值结果较差. 相反，JWAE 通过学习匹配联合分布从而插值出更高质量的视频帧. 从图 3 第二和第四行可知，其他方法受到插值结果的影响而导致翻译结果较差. 相反，JWAE 通过利用足够的跨域信息翻译出质量很好的视频帧.

第二，SYNTHIA 数据集上的定性分析如下：本文将冬季作为源域，首先在冬季域进行视频插值，然后把它们翻译到春夏秋季的域. 在图 4 中，本文同时比较了插值产生的视频帧与翻译的视频帧. 第一，当插值中间帧时，JWAE 能够生成更清晰的图像（参见图 4 的第一行）. 第二，JWAE 在春夏秋季三个域的翻译视频帧比其他对比方法更逼真（参见图 4 中的汽车形状）. 这些结果证明了 JWAE 可以有效地生成在视觉上高质量的视频序列.

第三，高帧率跨域视频合成如下：由于高帧率视频合成任务更具有挑战性，因此本文进一步研究不同方法在 Cityscapes 数据集上高帧率视频合成的性能. 与传统跨域视频合成不同，本文在源域中使用两个连续的视频帧作为输入，并插入中间 7 帧（即一个 8×帧率的上转换）. 在实验中，本文将语义分割域和街景域分别作为源域和目标域，用来合成跨域视频. 视频序列合成结果如图 5 所示.

由图 5 可得，本文首先比较了不同方法在源域中插值视频帧的质量（参见图 5 的最后一行）. 相比大多数的视频合成方法，JWAE 能够插值出质量更高的视频帧. 其次，本文还比较了目标域中的翻译帧（参见图 5 的中间行）. JWAE 能够在目标域中生成更清晰的图像. 最后，图 5 的首行可知 JWAE 合成的整个视频序列，它能够同时在两个域中生成连续的视频序列.

5.3 超参数 λ_z 的敏感度分析

为了探索公式(6)中超参数 λ_z 的取值对翻译性能的影响，本文在“冬季↔夏季”任务上比较了不同 λ_z 值[0.01, 0.1, 1, 10]对 FID 与 FID4Video 两种指标的影响，结果如表 3 所示. 当设置超参数 λ_z 的值为 0.1 时，JWAE 在两种指标上都达到最好的性能. 当逐渐增加 λ_z 的值到 1 和 10 时，模型性能逐渐下降. 当减少 λ_z 的值到 0.01 时也有同样的现象. 说明当 $\lambda_z = 0.1$ 时，JWAE 能够在潜在空间和数据空间的优化之间实现更好的权衡，从而获得更好的性能.

表 3 超参数 λ_z 的选取对翻译性能的影响. 在“冬季↔夏季”任务上比较 FID 和 FID4Video.

λ_z	冬季→夏季		夏季→冬季	
	FID	FID4Video	FID	FID4Video
0.01	94.91	20.29	107.65	18.90
0.1	**77.12**	**17.99**	**89.03**	**17.36**
1	89.07	21.04	102.18	18.63
10	101.07	23.66	108.47	20.50

表 4 不同方法在图像翻译和跨域视频合成任务中的运行时间对比.

“冬季→夏季”图像翻译任务测试时间					
方法	CoGAN	CycleGAN	AGGAN	UNIT	JWAE
时间/秒	0.128	0.207	0.206	0.107	0.168
“冬季→夏季”跨域视频合成任务测试时间					
方法	DVF-Cycle	AdaConv-Cycle	Slomo-Cycle	UNIT	JWAE
时间/秒	4.565	4.065	1.835	0.638	0.950
“冬季→夏季”任务训练时间					
方法	CoGAN	CycleGAN	AGGAN	UNIT	JWAE
时间/时	12.7	17.5	22.2	36.1	19.4

5.4 效率分析

在本小节中，本文比较了不同方法在图像翻译和跨域视频合成任务的运行时间. 在算法实现中，本文在一块英伟达 TITAN Xp GPU 上训练并测试不同方法在单张 SYNTHIA 数据集图片的平均运行时间. 由表 4 可知，在图像翻译任务中，不同方法的测试时间非常接近，更重要地，JWAE 翻译的图像质量高于其他方法（见表 1）. 在跨域视频合成任务中，JWAE 的测试时间次于 UNIT（相差在 1 秒之内），但 JWAE 生成的视频序列质量高于 UNIT（见图 3、图 4 和表 2）. 这两种方法的测试时间相比其他方法具有明显的优势，这是因为它们同时具备图像插值和翻译的功能，可以端到端地进行合成视频. 相反，另外几种对比方法分为插值和翻译两步进行，导致所损耗的时间大大增加. 对于训练时间，JWAE 在图像翻译任务中的训练时间次于 CoGAN 和 CycleGAN，但 JWAE 生成的图像质量远远高于这两种方法（见图 2 和表 1）.

6 结论

本文提出了一个新的联合 Wasserstein 自编码器（JWAE）模型来匹配域间的联合分布. 基于最优传输理论，JWAE 能够利用不同域之间的相关性来提高模型的翻译性能. 由于直接优化联合分布的 Wasserstein 距离的原始问题会带来巨大的计算代价，因此本文不直接优化该问题，而是基于最优传输理论推导出一个重要的定理，使得原问题转化成一个简单的优化问题. 本文在多个数据集上进行了大量无监督图像到图像翻译和跨域视频合成实验，实验结果均证明了本文提出的方法优于现有方法.

参 考 文 献

[1] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2223-2232.

[2] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks//Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 700-708.

[3] Bashkirova D, Usman B, Saenko K. Unsupervised video-to-video translation. arXiv preprint arXiv, 2018:1806.03698.

[4] Wang T C, Liu M Y, Zhu J Y, et al. Video-to-video synthesis//Proceedings of Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 1144--1156.

[5] Lindvall T. Lectures on the coupling method. United Kingdom: Courier Corporation, 2002.

[6] Kim T, Cha M, Kim H, et al. Learning to discover cross-domain relations with generative adversarial networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1857-1865.

[7] Yi Z, Zhang H, Tan P, et al. Dualgan: unsupervised dual learning for image-to-image translation//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 2223-2232.

[8] Li C, Liu H, et al. Alice: towards understanding adversarial learning for joint distribution matching//Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5495-5503.

[9] Pu Y, Dai S, Gan Z, et al. Jointgan: multi-domain joint distribution learning with generative adversarial nets//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4151-4160.

[10] Genevay A, Peyré~G, et al. Learning generative models with sinkhorn divergences//Proceedings of International Conference on Artificial Intelligence and Statistics. Long Beach, USA, 2018: 1608-1617.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of Advances in Neural Information Processing Systems. Montréal, Canada, 2014: 2672-2680.

[12] Cao J, Guo Y, Wu Q, et al. Adversarial learning with local coordinate coding//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4151-4160.

[13] Salimans T, Zhang H, Radford A, et al. Improving GANs using optimal transport//Proceedings of International Conference on Learning Representations. Vancouver, Canada, 2018.

[14] Kingma D P, Welling M. Auto-encoding variational bayes//Proceedings of International Conference on Learning

Representations. Banff, Canada, 2014.

[15] Tolstikhin I, Bousquet O, Gelly S, et al. Wasserstein auto-encoders//Proceedings of International Conference on Learning Representations. Vancouver, Canada, 2018.

[16] Guo, Y., Zheng, Y., Tan, M., Chen, Q., Chen, J., Zhao, P., & Huang, J. NAT: neural architecture transformer for accurate and compact architectures//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 735-747.

[17] Cao, J., Mo, L., Zhang, Y., Jia, K., Shen, C., Tan, M. Multi-marginal wasserstein gan//Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 1774-1784.

[18] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1125-1134.

[19] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans//Proceedings of the IEEE Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8798-8807.

[20] Li M, Huang H, Ma L, et al. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 184-199.

[21] Gokaslan A, Ramanujan V, et al. Improving shape deformation in unsupervised image-to-image translation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 649-665.

[22] Zhang R, Pfister T, Li J. Harmonic unpaired image-to-image translation//Proceedings of International Conference on Learning Representations. New Orleans, USA, 2019.

[23] Mejjati Y A, Richardt C, Tompkin J, et al. Unsupervised attention-guided image-to-image translation//Proceedings of Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 3693-3703.

[24] Liu Z, Yeh R A, Tang X, et al. Video frame synthesis using deep voxel flow//Proceedings of International Conference on Computer Vision. Venice, Italy, 2017: 4463-4471.

[25] Ji D, Kwon J, McFarland M, et al. Deep view morphing//Proceedings of the IEEE Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 2155-2163.

[26] Niklaus S, Mai L, Liu F. Video frame interpolation via adaptive separable convolution//Proceedings of International Conference on Computer Vision. Venice, Italy, 2017: 261-270.

[27] Jiang H, Sun D, Jampani V, et al. Super slomo: High quality estimation of multiple intermediate frames for video interpolation//Proceedings of the IEEE Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9000-9008.

[28] Liu M Y, Tuzel O. Coupled generative adversarial networks//Proceedings of Advances in Neural Information Processing Systems. Barcelona Spain, 2016: 469-477.

[29] Donahue J, Krahenbuhl P, Darrell T. Adversarial feature learning//Proceedings of International Conference on Learning Representations. Toulon, France, 2017.

[30] Dumoulin V, Belghazi I, Poole B, et al. Adversarially learned inference//Proceedings of International Conference on Learning Representations. Toulon, France, 2017.

[31] Villani C. Optimal transport: old and new. Springer: Springer Science & Business Media, 2008.

[32] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks//Proceedings of International Conference on Machine Learning. Sydney, Australia, 2017: 214-223.

[33] Lei N, Su K, et al. A geometric view of optimal transportation and generative model. Computer Aided Geometric Design, 2019, 68: 1-21.

[34] Peyré G, Cuturi M. Computational optimal transport. Foundations and Trends in Machine Learning, 2019, Vol. 11: 355-607

[35] Kingma D P, Ba J. Adam: a method for stochastic optimization//Proceedings of International Conference on Learning Representations. San Diego, CA, 2015

[36] Cordts M, Omran M, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223.

[37] Ros G, Sellart L, Materzynska J, et al. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes//Proceedings of the IEEE Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3234-3243.

[38] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans//Proceedings of Advances in neural information processing systems. 2016: 2234-2242.

[39] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE Computer Vision and Pattern Recognition. 2016: 2818-2826.

[40] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 6626-6637.

[41] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset//Proceedings of the IEEE Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 6299-6308.

附录 1.

补充材料可以在https://github.com/caojiezhang/paper/中获取.

附录 2.

定理 1. 给定解码器 G_1 和 G_2，并假设其确定性模型 $P_{G_1}(X'|Z)$ 和 $P_{G_2}(Y'|Z)$ 是狄拉克函数，定义为 $P_{G_1}(X'|Z=\mathbf{z})=\delta_{G_1(\mathbf{z})}$ 和 $P_{G_2}(Y'|Z=\mathbf{z})=\delta_{G_2(\mathbf{z})}$，对于 $\forall \mathbf{z}\in Z$，则 P_A 和 P_B 的 Wasserstein 距离可以定义为:

$$\begin{aligned}W_c(P_A,P_B)=&\inf_{Q\in Q_1}\mathrm{E}_{P_X}\mathrm{E}_{Q(Z_1|X)}\left[c_1\left(X,G_1(Z_1)\right)\right]\\&+\inf_{Q\in Q_2}\mathrm{E}_{P_Y}\mathrm{E}_{Q(Z_2|Y)}\left[c_2\left(G_2(Z_2),Y\right)\right],\end{aligned}\tag{11}$$

其中定义两个集合 $Q_1=\{Q(Z_1|X)|Q\in\tilde{Q},P_Y=Q_Y\}$ 和 $Q_2=\{Q(Z_2|Y)|Q\in\tilde{Q},P_X=Q_X\}$ 为所有编码器的集合，而 Q 属于 $\tilde{Q}=\{Q|P_{Z_1}=Q_{Z_1},P_{Z_2}=Q_{Z_2}\}$.

证明. 本文首先定义集合 $\mathrm{P}(X\sim P_X,X'\sim P_{G_1})$ 和 $\mathrm{P}(Y\sim P_Y,Y'\sim P_{G_2})$ 为 (X,X') 和 (Y,Y') 的所有联合分布的集合，其中它们的边缘分布分别为 P_X,P_{G_1} 和 P_Y,P_{G_2}. 定义 $\mathrm{P}(P_A,P_B)$ 为 P_A 和 P_B 的所有联合分布的集合. 根据 Wasserstein 距离 $W_c(P_A,P_B)$ 定义，本文有如下推导

$$W_c(P_A,P_B)\tag{12}$$

$$=\inf_{P\in\mathrm{P}(P_A,P_B)}\mathrm{E}_P[c(X,Y';X',Y)]\tag{13}$$

$$=\inf_{P\in\mathrm{P}(P_A,P_B)}\mathrm{E}_P[c_1(X',X)]+\inf_{P\in\mathrm{P}(P_A,P_B)}\mathrm{E}_P[c_2(Y',Y)]\tag{14}$$

$$=\inf_{P\in\mathrm{P}_{X,X'}}\mathrm{E}_P[c_1(X',X)]+\inf_{P\in\mathrm{P}_{Y,Y'}}\mathrm{E}_P[c_2(Y',Y)]\tag{15}$$

$$=\inf_{P\in\mathrm{P}(P_X,P_{G_1})}\mathrm{E}_P[c_1(X',X)]+\inf_{P\in\mathrm{P}(P_Y,P_{G_2})}\mathrm{E}_P[c_2(Y',Y)]\tag{16}$$

$$=W_{c_1}(P_X,P_{G_1})+W_{c_2}(P_{G_2},P_Y).\tag{17}$$

基于 Wasserstein 距离 $W_c(P_A,P_B)$ 的定义，公式(13)成立. 基于距离函数的定义，公式(14)成立 公式(15)成立是因为变量组合 (X,X') 和 (Y,Y') 互相独立，且由联合分布 $\mathrm{P}_{X,Y',X',Y}$ 可推出 $\mathrm{P}_{X,X'}$ 和 $\mathrm{P}_{Y,Y'}$ 分别是 (X,X') 和 (Y,Y') 上的边缘分布. 对于公式(16)，如果 $P_{G_1}(X'|Z)$ 和 $P_{G_2}(Y'|Z)$ 是狄拉克分布（满足 $X'=G_1(Z)$ 和 $Y'=G_2(Z)$），则 $\mathrm{P}_{X,X'}=\mathrm{P}(P_X,P_{G_1})$ 和 $\mathrm{P}_{Y,Y'}=\mathrm{P}(P_Y,P_{G_2})$ 成立.

本文考虑关于变量 $(X,X',Z_1)\in\mathrm{X}\times\mathrm{X}\times\mathrm{Z}$ 的联合分布集 P_{X,X',Z_1} 以及关于变量 $(Y,Y',Z_2)\in\mathrm{Y}\times\mathrm{Y}\times\mathrm{Z}$ 的联合分布集 P_{Y,Y',Z_2}. 本文分别定义

$$\mathrm{P}(X\sim P_X,Z_1\sim P_{Z_1})\text{和}\mathrm{P}(Y\sim P_Y,Z_2\sim P_{Z_2})\tag{18}$$

为 (X,Z_1) 和 (Y,Z_2) 的所有联合分布集，且边缘分布分别为 P_X,P_{Z_1} 和 P_Y,P_{Z_2}. P_{X,X',Z_1} 满足 $X\sim P_X,(X',Z_1)\sim P_{G_1,Z_1}$ 和 $(X'\perp X)|Z_1$. 对于 P_{Y,Y',Z_2} 同样如此. 由 P_{X,X',Z_1} 的定义，本文可以分别定义 $\mathrm{P}_{X,X'}$ 和 P_{X,Z_1} 为 (X,X') 和 X,Z_1 上的边缘分布集，同样地可以定义 $\mathrm{P}_{Y,Y'}$ 和 P_{Y,Z_2}.

$$\begin{aligned}&W_{c_1}(P_X,P_{G_1})+W_{c_2}(P_{G_2},P_Y)\\&=\inf_{P\in\mathrm{P}_{X,X',Z_1}}\mathrm{E}_P[c_1(X',X)]+\inf_{P\in\mathrm{P}_{Y,Y',Z_2}}\mathrm{E}_P[c_2(Y',Y)]\end{aligned}\tag{19}$$

$$\begin{aligned}&=\inf_{P\in\mathrm{P}_{X,X',Z_1}}\mathrm{E}_{P_{Z_1}}\mathrm{E}_{X\in P(X|Z_1)}\mathrm{E}_{X'\in P(X'|Z_1)}[c_1(X',X)]\\&\quad+\inf_{P\in\mathrm{P}_{Y,Y',Z_2}}\mathrm{E}_{P_{Z_2}}\mathrm{E}_{Y\in P(Y|Z_2)}\mathrm{E}_{Y'\in P(Y'|Z_2)}[c_2(Y',Y)]\end{aligned}\tag{20}$$

$$\begin{aligned}&=\inf_{P\in\mathrm{P}_{X,X',Z_1}}\mathrm{E}_{P_{Z_1}}\mathrm{E}_{X\in P(X|Z_1)}[c_1(X,G_1(Z_1))]\\&\quad+\inf_{P\in\mathrm{P}_{Y,Y',Z_2}}\mathrm{E}_{P_{Z_2}}\mathrm{E}_{Y\in P(Y|Z_2)}[c_2(G_2(Z_2),Y)]\end{aligned}\tag{21}$$

$$\begin{aligned}&=\inf_{P\in\mathrm{P}_{X,Z_1}}\mathrm{E}_{(X,Z_1)\in P}[c_1(X,G_1(Z_1))]\\&\quad+\inf_{P\in\mathrm{P}_{Y,Z_2}}\mathrm{E}_{(Y,Z_2)\in P}[c_2(G_2(Z_2),Y)]\end{aligned}\tag{22}$$

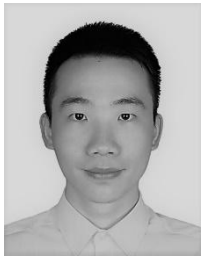
$$\begin{aligned}&=\inf_{P\in\mathrm{P}(X,Z_1)}\mathrm{E}_{(X,Z_1)\in P}[c_1(X,G_1(Z_1))]\\&\quad+\inf_{P\in\mathrm{P}(Y,Z_2)}\mathrm{E}_{(Y,Z_2)\in P}[c_2(G_2(Z_2),Y)]\end{aligned}\tag{23}$$

$$\begin{aligned}&=\inf_{Q\in Q_1}\mathrm{E}_{P_X}\mathrm{E}_{Q(Z_1|X)}[c_1(X,G_1(Z_1))]\\&\quad+\inf_{Q\in Q_2}\mathrm{E}_{P_Y}\mathrm{E}_{Q(Z_2|Y)}[c_2(G_2(Z_2),Y)],\end{aligned}\tag{24}$$

其中本文定义集合 $Q_1=\{Q(Z_1|X)|Q\in\tilde{Q},P_Y=Q_Y\}$ 和集合 $Q_2=\{Q(Z_2|Y)|Q\in\tilde{Q},P_X=Q_X\}$ 是所有概率编码器的集合，而 Q_1 和 Q_2 分别表示为 $Z_1\sim Q(Z_1|X)$ 和 $Z_2\sim Q(Z_2|Y)$ 的边缘分布.

公式(19)使用了全期望公式，根据 P_{X,X',Z_1} 条件独立的性质可得公式(20). 公式(22)计算了关于 X' 和 Y' 的期望，然后用全部概率表示. 公式(23)利用了 $\mathrm{P}_{X,Z_1}=\mathrm{P}(X\sim P_X,Z_1\sim P_{Z_1})$ 和 $\mathrm{P}_{Y,Z_2}=\mathrm{P}(Y\sim P_Y,Z_2\sim P_{Z_2})$，因为 $\mathrm{P}(P_X,P_{G_1}),\mathrm{P}_{X,X',Z_1}$ 和 $\mathrm{P}_{X,Y}$ 取决于条件分布 $\mathrm{P}_{G_1}(X'|Z_1)$ 的选择，而 P_{X,Z_1} 不需要. 对 Y 和 G_2 的分布也是同样如此. 公式(24)做了一个变换. 当 $Q_{Z_1}=Q_{Z_2}$ 和 $P_Y=Q_Y$ 时，生成模型 $Q(Z_1|X)$ 由两个情况推导出，即 Z_1 从 $E_1(X)$ 和 $E_2(G_2(E_1(X)))$ 获取. 另外，生成模型 $Q(Z_2|Y)$ 同样得出.

证毕.



CAO Jie-Zhang, M.S. candidate. His research interest focuses on adversarial machine learning.

MO Lang-Yuan, M.S. candidate. His research interest focuses on computer vision.

DU Qing, Ph.D., Associate Professor. His research interest focuses on deep learning.

GUO Yong, Ph.D. candidate. His research interest focuses on deep learning.

ZHAO Pei-Lin, Ph.D. His research interests include machine learning and data mining.

HUANG Jun-Zhou, Ph.D. Associate Professor, Ph.D. supervisor. His research interests include machine learning, computer vision.

TAN Ming-Kui, Ph.D. professor， Ph.D. supervisor. His research interest focuses on machine learning.

Background

Joint distribution matching (JDM) seeks to learn the bidirectional mappings to match the joint distributions of unpaired data in two different domains. This problem has been applied in computer vision, such as image translation[1,2]and video synthesis[3,4]. However, learning the joint distribution of two domains is more difficult and has two key challenges.

The first key challenge, from a probabilistic modeling perspective, is how to exploit the joint distribution of unpaired data by learning the bidirectional mappings between two different domains. In the unsupervised learning setting, there are two sets of samples drawn separately from two marginal distributions in two domains. Based on the coupling theory[5], there exist an infinite set of joint distributions given two marginal distributions, and thus infinite bidirectional mappings between two different domains may exist. Therefore, directly learning the joint distribution without additional information between the marginal distributions is a highly ill-posed problem. Recently, many studies[1,6,7] have been proposed to learn the mappings in two domains separately, which cannot learn cross-domain correlations. Therefore, how to exploit sufficient information from the joint distribution remains an open question.

The second critical challenge is how to formulate and optimize the joint distribution matching problem. Most existing methods[8,9] do not directly measure the distance between joint distributions, which may result in the distribution mismatching issue. To address this, one can directly apply some statistics divergence, *e.g.*, Wasserstein distance, to measure the divergence of joint distributions. However, the optimization may result in intractable computational cost and statistical difficulties[10]. Therefore, it is important to design a new objective function and an effective optimization method for the joint distribution matching problem.

In this paper, we propose a Joint Wasserstein Auto-Encoder (JWAE) method. Specifically, to address the first challenge, we use optimal transport theory to exploit geometry information and correlations between different domains. For the second challenge, we apply Wasserstein distance to measure the divergence between joint distributions in two domains, and optimize it based on an equivalence theorem.

The contributions are summarized as follows:

i) Relying on optimal transport theory, we propose a novel JWAE to solve the joint distribution matching problem. The proposed method is able to exploit sufficient information for JDM by learning correlations between domains instead of learning from individual domain.

ii) We derive an important theorem so that the intractable primal problem of minimizing Wasserstein distance between joint distributions can be readily reduced to a simple optimization problem.

iii) We apply JWAE to unsupervised image translation and cross-domain video synthesis. Experiments on real-world datasets show the superiority of the proposed method over several state-of-the-arts.

This work was partially supported by Guangdong Provincial Scientific and Technological Funds under Grants 2018B010107001, National Natural Science Foundation of China (NSFC) 61836003 (key project), Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183, Fundamental Research Funds for the Central Universities D2191240, Tencent AI Lab Rhino-Bird Focused Research Program (No. JR201902).