分类号 G623.58 密级 公开

UDC 004.8

学 位 论 文

基于图像特征的人脸编辑研究

（题名和副题名）

严元杰

（作者姓名）

指导教师姓名、职务、职称、学位、单位名称及地址 申富饶教授

南京大学计算机科学与技术系 江苏省南京市栖霞区仙林大道 163 号 210023

申请学位级别 博士 专业名称 计算机科学与技术

论文提交日期 2023 年 3 月 25 日 论文答辩日期 2023 年 8 月 21 日

学位授予单位和日期

答辩委员会主席： 杨明教授

评阅人： 魏秀参教授

武港山教授

戴新宇教授

路通教授

南京大学

研 究 生 毕 业 论 文

（申 请 博 士 学 位）

论　文　题　目　基于图像特征的人脸编辑研究

作　者　姓　名　严元杰

学科、专业名称　计算机科学与技术

研　究　方　向　神经网络及计算机视觉

指　导　教　师　申富饶教授

2023 年 **11** 月 **27** 日

学　　号：**DZ1833030**

论文答辩日期：**2023** 年 **8** 月 **21** 日

指 导 教 师：　　（签字）

Research about Face Editing Based on Image Features

by

Yan Yuanjie

Supervised by

Professor Shen Furao

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Computer Science and Technology

Department of Computer Science and Technology
Nanjing University

November 27, 2023

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于图像特征的人脸编辑研究

计算机科学与技术 专业 2018 级博士生姓名：严元杰

指导教师（姓名、职称）：申富饶教授

摘　要

人脸编辑围绕如何对人脸图像进行改变或修饰等问题展开。目前，人脸编辑的研究非常活跃，涉及的领域包括计算机视觉、三维模型和机器学习等。随着人脸编辑的深入研究，基于神经网络的方法在该问题上取得了巨大进展。人脸特征成为利用神经网络解决人脸编辑的关键研究内容。第一，不同姿态的人脸需要对齐到标准姿态的人脸，有助于减轻人脸编辑的难度进而提升生成图像的质量。人脸图像对齐作为预处理过程，可用于统一人脸特征编解码过程。第二，为了实现人脸语义内容的编辑，需要利用网络模型提取有效的人脸特征编码。第三，基于人脸图像特征的生成研究是人脸编辑的关键步骤，实现了逼真人脸的生成。第四，综合上述的研究成果，实现了基于特征编码的人脸图像编辑。依据人脸编辑问题的解决过程，本文将研究内容划分为关于人脸图像的局部点云对齐、人脸图像的特征编码、基于特征编码的人脸图像生成和基于特征编码的人脸编辑。具体的研究内容与贡献如下：

1. 利用局部点云数据，部分解决人脸对齐问题。人脸编辑主要针对标准姿态的人脸进行处理。对于任意获取的人脸图像，将其对齐到标准姿态的预处理步骤，方便后续人脸图像的编解码过程。我们研究了成对局部点云的对齐及补全任务，并提出了基于混合优化的点云对齐方法。该方法利用无约束变量在整个变换矩阵空间进行迭代优化，解决了点云的局部和全局匹配，提高了对齐精度。该研究可以辅助进行人脸对齐，为后续人脸编辑进行铺垫。
2. 部分解决了无标注人脸图像的特征编码问题。目前，学习具有强泛化性的图像特征编码依赖于海量的标注数据。然而，在图像特征相关的任务中，标注海量数据是十分困难甚至是不可行的。本文立足于无标注的图像数据，结合图像聚类任务，提出了新颖的聚类方法，能够有效地学习图像特

征编码。我们的方法通过引入无监督高斯混合聚类，提升了人脸语义特征编码的学习效果。该研究实现了人脸编辑问题中无标注图像的特征编码的学习。

3. 分析了 StyleGAN 中 Style 特征解码的图像生成过程，研究了不同学习方法在图像特征解码时作用，并着重研究了人脸特征解码的生成过程。虽然基于特征的图像生成模型具有良好的生成效果，但缺乏对于特征在网络中解码过程的深入研究。基于 StyleGAN 的图像生成模型，分析了该预训练的模型中的 Style 特征的解码过程。基于 Style 编码的图像生成模型增强了特征编码空间到图像空间的可解释性，提高了生成图像的质量。该研究分析了基于 Style 编码的合成网络在人脸生成时的解码过程，明确了特征编码空间和人脸图像语义空间的对应关系。
4. 实现了指定语义内容的特征通道定位，并完成了连续可控的人脸图像编辑。在图像编解码研究的基础上，本文探索了特征编码和语义内容之间的关联，在编码隐空间上实现了对指定语义内容的编辑。针对人脸属性编辑任务，在 StyleGAN 模型的隐编码空间 $\boldsymbol{S}$ 上定位与指定属性有关的特征通道。最后，还提出对特征编码的单通道和多通道编辑方法，细粒度地控制图像的生成内容。该研究利用合成网络的分层 Style 编码提出了可指定人脸属性的编辑方法，通过结合人脸图像特征编解码过程，实现了多种人脸属性编辑。

综上所述，本文围绕人脸图像对齐、编码、解码和编辑这四个问题进行了递进式地深入研究。相关实验及分析验证了本文在关于人脸图像编辑研究方面的前沿性，为研究新的图像编辑问题提供了有价值的参考。

关键词： 人脸图像; 图像特征编码; 图像特征解码; 局部点云对齐; 图像特征编辑

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research about Face Editing Based on Image Features

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Yan Yuanjie

MENTOR: Professor Shen Furao

Abstract

Facial editing revolves around changing or modifying facial attributes. At present, research on face editing is very active, involving computer vision, 3D models, machine learning and other fields. As face editing has been intensively studied, network-based methods have made great progress on this issue. Face features have become the key research content of using neural networks to solve face editing. First, the faces of different poses should be aligned with the faces of standard poses to reduce the difficulty of face editing and improve the effect of generated images. We need to align face images to make a unified encoding and decoding process of face features. Second, to achieve the editing of the semantic content of the face, it is necessary to extract face feature encoding via the network model. Third, image generation based on face image features is a key step in face editing, which generates realistic faces. Fourth, we study the problem of image editing based on facial feature encoding. According to the process of the face editing problem, this paper divides the research content into feature encoding of face images, face image generation based on feature encoding, point cloud alignment of face models and face editing based on feature encoding. The specific research contents and contributions are as follows:

1. We have solved the problem of feature encoding learning under unlabelled image data. Learning image features with great generalization relies on immense amounts of annotated data. However, annotating data is difficult or even infeasible in some image tasks. Based on the unlabelled image, this paper focuses on image clustering which studies image feature encoding with unsupervised learning. We propose the Deep Embedded Dimensionality Reduction Clustering (DERC) method. This study initially investigates the feature encoding of images in face editing.

2. We analyse the image generation process of style features in StyleGAN. We focus on image generative models of StyleGAN. This paper analyses the decoding process of the style features in the pre-trained model. The style features improve the interpretability from coding space to image space and the quality of generated images. This study preliminarily analyses the decoding process of the Synthetic Network based on Style encoding when generating faces.
3. We study the alignment of local point clouds. Face editing mainly deals with faces in standard poses. For any acquired face image, the preprocessing step of aligning it to the standard pose is beneficial to the encoding and decoding process of the face image. We study the alignment and completion tasks of paired local point clouds and propose a hybrid optimization-based method for point cloud alignment. The method uses unconstrained variables to perform iterative optimization in the entire transformation matrix space, which solves the local and global matching of point clouds. Face alignment can pave the way for subsequent face editing.
4. We achieve the feature channel location of the specified semantic content and conduct fine-grained image editing. Based on the research on image feature encoding and decoding, this paper investigates the mapping between image encoding and semantic content. We implemented the specified semantic content manipulation on the latent feature space. On the face attribute editing, we detect the feature space $\boldsymbol{S}$ in StyleGAN to locate the channel of the specified attribute. We also propose single-channel and multi-channel editing methods to control the fine-grained content of images. This study proposes an editing method that can specify face attributes by using the hierarchical Style feature of the synthesis network. We accomplish various face attribute editing based on the encoding and decoding research of face image features.

To sum up, this paper conducts progressive research on the four issues of encoding, decoding, alignment and editing of face images. Relevant experiments and analysis prove that this paper is at the forefront of research on face image editing, and provides a valuable reference for the study of new image editing issues.

keywords: Face Image; Image Feature Encoding; Image Feature Decoding; Point Cloud Alignment; Image Feature Editing

目　　录

插图清单

附表清单

第一章　绪论

1.1　研究背景及意义

图像编辑是指使用计算机软件对数字图像进行修改、增强和优化的过程。传统的编辑修改内容包括调整图像的亮度、对比度、色彩平衡、去除噪点、旋转缩放、添加文字等。然而，这些编辑内容往往不涉及图像语义内容的修改。目前，深度学习在编辑图像语义内容呈现了巨大的研究潜力。其中，以人脸图像作为主要编辑内容的研究不断涌现。例如，人脸替换技术实现了对图像中全局人脸的替换，并实现了虚假视频的伪造。人脸动漫化技术能够将任意人脸进行动漫化，依据真实人脸生成相似的虚拟动漫头像。人脸属性编辑技术能够细粒度地控制人脸中不同语义内容的生成。这些关于人脸图像的编辑技术在动漫设计，视频生产和社交娱乐上具有重要的应用价值。由于人脸编辑属于图像编辑的一条重要分支问题，本文主要围绕图像编辑问题展开研究，并着重实现面向人脸图像属性内容的编辑。

同时，面向不同目标的图像编辑任务被不断地提出并加以研究。新提出的各种的图像编辑方法大都围绕图像特征展开。首先，图像特征可以看作为高维图像映射到低维的特征编码。特征编码也可以表示图像特征的提取过程，用于提取图像中的语义属性。基于深度学习的图像编辑方法通常将特征编码作为其关键的研究内容。其次，依据解码目标不同，特征编码可被用于多种图像任务。在图像识别中，图像特征解码可以看作为对特征编码的判别问题。在目标检测中，图像特征解码可以看作利用特征编码的回归问题。将图像特征解码到原始的图像也称为基于特征编码的图像生成问题。本文主要研究基于特征编码的生成方法，并用于解决图像编辑问题。为表述简便，本文采用图像特征解码代指基于特征编码的图像生成问题。接着，图像特征编码和解码实现了图像域到特征域的双向映射。为研究特征空间与图像空间更深层次的对应关系，图像特征编辑面向可控的图像生成问题，其旨在揭示特征编码与语义内容的关联。这些工作丰富了图像编辑的研究内容，也赋予其新的研究意义。因此，人脸编辑问题的解决过程离不开对图像特征的研究。

在众多的图像编辑问题及其应用场景中，虽然所处理的内容大相径庭，但都涉及到对图像特征的研究。本文围绕人脸图像编辑问题，按照递进式的解决思路，将研究内容划分为四个研究层次。首先，研究 3D 点云的对齐问题，其可作为人脸图像预处理操作，统一人脸图像特征编解码的过程。为了实现合理的人脸编辑，需要对输入的人脸进行裁剪，并使其与标注姿态人脸进行对齐。由于人脸等物体在二维图像中容易收到拍摄角度，人脸表情等影响，难以生成具有稳定表示的特征。我们将研究三维点云模型的对齐任务。该研究方便端到端地实现人脸到特征，特征到人脸的映射。其次，研究图像特征编码问题，学习人脸图像到特征编码的表达。深度网络模型需要有效地学习图像特征，这离不开对模型的类型和结构的探索。卷积操作能提取图像相邻区域内像素的特征。通过逐层搭建的卷积层，深度神经模型能提取高级的语义信息。因此，卷积神经网络（Convolutional Neural Networks，CNN）被广泛地应用于图像任务中，例如 ResNet 分类模型 [1]，全卷积网络 FCN 语义分割模型 [2] 等。卷积神经网络成为了提取图像特征的主流模型。最近，研究者又提出了基于 Transformer 的图像特征提取网络，试图挑战 CNN 在图像处理领域的统治地位。ViT[3] 模型将图像看作为图像块序列进行处理，利用 Transformer 捕获了全局的图像特征。这使得 ViT 可以超越卷积神经网络模型的分类精度。这引发了采用 Transformer 提取图像特征，并应用于下游不同图像任务的浪潮。然而，对于不同的网络模型，如何训练其学习具有良好泛化性能的图像特征仍是十分重要的研究内容。之后，研究图像特征解码问题，有效利用人脸特征生成逼真的人脸图像。网络模型需要合理地利用图像特征，才能解决各类图像问题。因此，需要对图像特征进行有效的解码处理以满足不同任务的需求。在图像分类中，可以构建神经网络分类器，对图像特征进行分类。在语义分割和图像生成等任务中，需要对图像特征进行更加复杂的处理。其中，基于图像特征的生成是研究的难点。本文将聚焦于图像特征解码中基于图像特征的生成问题的研究。基于图像特征的生成是研究将低维的图像特征映射回原始的图像空间的过程。我们需要研究图像特征空间与原图像空间的之间的映射关系，并实现图像和特征编码的对齐。最后，研究图像特征编辑问题，提出新的人脸图像编辑方法。理解图像特征并探索图像特征与图像语义的对应关系是解释深度网络模型的重要环节。传统手工提取的图像特征具有良好的可解释性，如角点特征和 SIFT 特征等。然而，由于神经网络中复杂的非线性映射常被作为黑盒模型，解释网络模型学习到的图像特征的语义信息是十分困难但有价值的研究问题。在

原始的图像域中，我们可以清楚地区分不同的语义内容，例如前景和背景，不同目标实体等。研究图像特征域中是否含有潜在的语义内容是理解图像特征的关键步骤。例如，图像特征的稀疏表示致力于学习到相互独立的解耦特征，用以解释图像特征的语义内容。基于上述的研究，图像特征编辑试图通过控制特征编码完成指定图像内容的修改。

另一方面，深度学习在计算机视觉的众多复杂场景任务中取得突破性进展，得益于海量的标注数据。在深度神经网络模型的性能大都受限于手工标注的数据样本。在机器学习中，当训练数据不足时，模型的参数估计会引起臭名昭著的过度拟合问题。由于神经网络模型含有数百万甚至数十亿的参数，这使得训练模型参数成为棘手的问题。目前，标记大型训练数据集是克服过拟合问题的有效解决方法。在计算机视觉中，ImageNet[4] 和 COCO[5] 被公开发布用于图像分类、语义分割、实例分割等问题。然而，从某种程度上，对于本文研究的基于深度学习的人脸图像编辑问题，无法实现手工标注数据。例如，要实现人脸属性的编辑，但无法明确修改后的人脸内容。这为神经网络模型学习造成了极大困难。因此，神经网络模型从未标记数据或少量标记样本中进行学习是当前研究的热点问题。对于无标记的数据，自监督学习 [6] 尝试利用辅助任务，从大规模的无标注数据中挖掘自身的监督信息，并以此构造待优化的损失函数，实现对神经网络模型的训练。对于少量标记的样本，小样本学习 [7] 旨在通过少数具有标注信息的训练样本，实现模型对未标记数据的推理并用于新阶段的学习。迁移学习 [8] 表明运用预训练模型已有的知识来学习新的知识，寻找到已有知识和新知识之间的联系，是解决新的相似问题的有效途径。当在少量标注样本上训练时，迁移学习能有效提高模型的泛化能力。同时，在利用神经网络处理图像的任务中，图像分类任务占有举足轻重的地位。在图像分类问题上提出的各种网络模块和结构，例如 ResNet 模块，多个小卷积核串联和跳层连接等，被广泛应用于其他问题的模型设计中。神经网络模型不仅被用于分类或回归等图像任务，也被用于图像生成任务。生成对抗网络（Generative Adversarial Network，GAN）[9] 引入了对抗的学习机制来生成伪造的样本。GAN 利用模型间的标注信息实现模型间的训练，弥补了只基于标注数据的学习方法的不足。这些学习方法的研究被广泛应用于上述神经网络的训练中。围绕图像特征的工作离不开对于这些学习方法的研究。因此，在人脸编辑的解决过程中，需要全面地分析神经网络的各种学习机制，利用图像特征完成指定属性内容的修改。

1.2 研究现状及难点

图像编辑是一个十分困难的问题，往往涉及多个不同的图像任务。按照递进式的研究路线，将其划分为图像的特征编码、基于图像特征的解码、基于局部点云特征的图像对齐和图像特征编辑这四大块研究内容。这些研究内容都离不开对图像特征的研究。在众多的图像处理子任务中，图像特征的研究也涉及到不同的方面。例如，图像分类任务试图提取图像语义特征用于目标识别。图像生成任务试图利用图像特征（或随机噪声）生成真实的伪造图像。图像特征编辑则被广泛应用于图像的二次编辑任务中，例如图像翻译和风格迁移。对于上述不同的图像任务，为清晰描述围绕图像特征的研究脉络，本文将图像特征的研究划分为图像特征编码、特征解码、特征对齐和特征编辑四个研究子问题。

为了有效地解决图像编辑问题，上述的研究任务离不开对神经网络学习机制的探索。图像编辑的研究难以被精确的定义在监督学习和无监督学习的学习框架下。根据不同的图像编辑任务，有些像素风格化的图像编辑研究可以采用传统的卷积滤波算法，用以实现不同的滤镜风格。有些图像语义内容的编辑需要结合图像语义分割等研究任务，采用海量的语义标注数据集，监督地训练神经网络实现图像语义的编辑。由于编辑图像没有明确的目标，更多的工作集中在无监督学习神经网络，使其完成各种内容的图像编辑效果。除此之外，机器学习中的无监督和有监督学习方法难以揭示围绕图像编辑不同研究内容的差异和联系。例如，基于标注数据集学习的图像特征和基于数据文本对学习的图像特征，都属于监督学习的范畴，但其学习的方式存在显著区别。从标注数据角度，基于数据集的方法大都需要手工进行标注，但基于数据文本对的方法采用图像与文本间的非人为标注的标签。数据的不同导致了目标任务的差异，也造成了网络模型和学习方法的差异。但上述的方法都面向学习图像特征编码，彼此间存在关联。此外，各种新学习方法被用于特定的研究问题，难以涵盖围绕图像特征的研究内容。例如，自监督学习和对比学习能够解决图像特征的学习问题。生成对抗学习可用于解决基于特征编码的真实感图像生成任务。不同学习方法彼此相互独立，但又存在些许联系。比较不同学习方法的差异和联系是个具有挑战性的问题。本文的图像编辑中所讨论的方法往往涵盖了多种学习方法。

本文的研究工作主要围绕人脸编辑问题展开。首先，详细论述了人脸编辑

的研究内容，并着重形式化定义了人脸属性编辑问题。其次，本节将从与人脸编辑相关的研究任务和学习方法两方面的内容展开，并介绍相关的研究现状。

1.2.1 人脸编辑的研究现状

广义上来说，本文主要面向解决图像编辑问题。其中，由于人脸在众多图像中的研究及应用中占有重要地位，如人脸动漫化和换脸等应用。人脸编辑可以看作为一类重要的图像编辑问题。本文围绕图像编辑的研究大都围绕人脸编辑问题展开。并且，相对任意图像内容而言，人脸图像包含的内容相对简单，其所涉及的相关技术相对成熟。例如，利用人脸定位技术获取单张的居中人脸，这些预处理技术有利于提高不同人脸编辑时的泛化性能。

同时，人脸编辑涵盖众多的人脸图像处理问题。传统的图像编辑技术，如模糊、锐化和色彩直方图等可直接应用于人脸图像。例如，利用高斯模糊实现人脸的磨皮，利用锐化实现人脸局部五官的调整。然而，这些编辑技术大都是面向图像中像素的色彩或光照调整。随着深度学习的发展，研究人员逐渐提出了各种面向人脸语义属性编辑的研究问题。人脸中含有丰富的语义属性。从五官角度出发，任意人脸可以看作为具有不同属性五官的组合。通过修改五官语义内容可以实现人脸修改。例如，仅头发而言，就有对发型、发色和发量等不同内容的编辑研究。从人脸表情出发，人脸能够表达不同的神情，如高兴、生气和伤心等。这种难以明确定义的语义表情丰富了人脸编辑问题的研究内容，但对其也提出新的挑战与困难。不同于人脸中像素，五官区域等语义内容直接联系的编辑问题研究，人脸表情或情感的语义研究是难以明确从图像的内容中反映的。

并且，由于深度学习大都采用端到端的解决方法，人脸编辑的研究大都面向可控语义内容生成的图像生成。在图像处理中，网络模型能过提取具有不同语义属性的人脸特征。在图像生成中，网络模型能够随机生成各种逼真的高清人脸。在人脸编辑的处理中，大都工作都采用结合上述研究内容的思路。首先，将原始人脸图像空间映射到语义特征空间。其次，结合明确的编辑内容在语义特征空间进行修改。最后，利用图像生成过程，将编辑后的特征进行解码生成修改后的人脸。为了进一步明确图像编辑问题的研究内容，本文将围绕语义属性展开人脸编辑的研究。人脸属性编辑针对人脸图像中丰富的属性内容，并实现对其内容的可控修改或生成。由于，本文主要从深度网络模型对其展开研究。将网络模型看作黑盒模型，人脸属性编辑将输入原始图像和指定的修改

语义内容，输出按照指定语义修改的生成图像，实现原图像域到目标图像域的可控修改。

人脸属性编辑问题的形式化定义：设原始的人脸图像为 X。根据预训练的 k 个属性二分类器模型集合 $\mathcal{F} = \{F_0, F_1, ..., F_k\}$，该图像 X 包含有 k 个不同的属性集合 $\mathcal{A} = \{a_0, a_1, ..., a_k\}$。其中，$a_i$ 可以看作为是否包含该属性的几率，取值为 $[0, 1]$。并设待修改的语义属性目标为 a_t，且 t 为第 t 个属性特征向量的下标。利用深度网络模型 M，修改后的编辑图像为 X'。人脸属性编辑的过程可以看作为结合指定语义内容 a_t，实现原图像域到目标图像域的生成过程。我们将其形式化为

$$X' = M(X, a_t). \tag{1-1}$$

当选取 a_t 进行内容编辑时，人脸属性编辑的优化目标是尽可能地只修改语义内容 a_t，而不改变其余属性的内容。设编辑后的图像的属性集合 $\mathcal{A} = \{a'_0, a'_1, ..., a'_k\}$，人脸属性编辑的优化目标需要同时考虑两种语义内容的修改，形式化表达为

$$X' = \arg\min_{X'}\Big(\sum_{F\in\mathcal{F}\setminus F_t}(F(X') - F(X))^2 - (F_t(X') - F_t(X))^2\Big). \tag{1-2}$$

其中，最小化 $\sum_{F\in\mathcal{F}\setminus F_t}(F(X') - F(X))^2$ 主要用于保持编辑后图像其余内容与原人脸内容的一致，使得改动尽可能对其余的语义属性没有影响。最小化 $-F(X))^2 - (F_t(X') - F_t(X))^2$ 即最大化 $F(X))^2 - (F_t(X') - F_t(X))^2$，其负责指定语义属性 a_t 的内容修改。

1.2.2 人脸编辑涉及研究内容的难点

本文面向人脸属性编辑问题，围绕图像特征将研究内容划分为四个子问题，分别为图像的特征编码、基于图像特征的解码、基于局部点云特征的图像对齐和图像特征编辑。首先，利用深度学习模型提取图像的特征编码，即实现了图像域到特征域的非线性映射，完成对图像信息的抽取。其次，又研究了在图像特征解码中重要的一类图像生成问题，即研究了特征域到图像域的非线性映射，完成图像特征编码到语义内容的变换。本文所研究的图像特征编码和解码可以看作为一对对偶问题，彼此之间存在高度联系。最后，基于上述的研究内容，本文探索了图像域内的语义内容与特征域内的特征向量之间的对应关系，并实现了特征编码空间上的语义内容编辑。因此，本节分别回顾了图像特

征编码、图像特征解码、基于局部点云特征的图像对齐和图像特征编辑的研究现状。

图像特征编码：图像特征编码是从图像中获取相对稳定的特征表达，例如颜色，纹理和目标物体等。其中，颜色，纹理和光照等容易表达的信息，称为低级的图像特征。图像中物体和语义内容等难以精确描述的信息，称为高级的语义特征。传统手工提取的图像特征主要试图解决图像的平移不变性、旋转不变性和缩放不变性等问题。手工设计的特征表述子能有效提取低级的图像特征，但难以处理图像的高级语义特征。并且，手工提取图像特征由于与实际的任务分离，需要人为对图像特征进行筛选和分析。深度学习利用图像特征编码器，试图从任务中学习图像的特征编码，用于端到端地解决图像问题。在早期利用神经网络解决计算机视觉问题的研究中，图像编码的研究是隐式地包含在具体的图像任务中的。由于图像识别任务与特征编码的研究关联密切，其成为在神经网络模型研究图像编码的最主要的研究任务。例如，预训练的分类卷积神经网络如 VGG[10]，ResNet 网络等都可以看作为图像的语义特征提取网络和全连接的分类网络的组合。迁移学习的研究表明基于卷积神经网络的特征编码器可以作为通用的图像特征提取器用于相关的任务中。基于图像识别任务的预训练图像特征编码器被广泛用于其他图像处理任务中。例如，目标检测和语义分割网络都可以采用预训练的特征提取器来搭建下游任务模型。将图像识别任务中的网络模块作为通用的图像特征提取器极大地简化了深度学习在图像上的特征提取的工作。然而，由于图像分类与其他图像问题的差异，这种图像特征编码与任务绑定的方式显然限制了特征编码在不同任务上的应用性能。

利用图像识别任务学习通用特征编码器通常面临三个主要问题。首先，所获取的特征提取器是面向图像识别任务的，这受限于图像分类的识别精度。研究者从模型架构角度出发，提出了一系列的改进方案。基于神经网络搜索的 EffieientNet 网络 [11] 超过了手工设计的网络结构。Transformer 网络 ViT[3] 也刷新了基于卷积神经网络的分类精度。图像分类任务的识别精度不断提高，隐式地提高了图像特征提取器的编码能力。其次，基于图像分类的图像特征编码器需要海量的标注数据。大型的标注数据集如 ImagNet，COCO[5] 等的提出，助力于图像问题的研究，使得可以学习表达能力更强的特征编码。研究表明 [12] 即使在标注了上百万张图片的 ImageNet 数据集上，大规模的网络模型仍无法得到充分地学习。显然，标注数据的增长跟不上模型规模的增长。从无标注数据中学习图像特征成为了研究趋势。受自然语言预训练模型的启发，CLIP

方法 [13] 试图从海量的图像文本对中学习特征编码，主要利用预训练的文本特征对齐图像特征。由于图像文本对的收集来自互联网数据且无须人为标注，这解决了海量标注数据的问题。然而，利用无标注数据增加了学习特征编码的难度，这需要研究新的学习机制。最后，不依赖于图像识别任务，直接面向图像特征进行学习是具有客观前景的。许多通用的图像特征提取器都依赖于在图像识别任务中进行学习。然而，这些研究的重点在于分类任务的精度，而不是图像的特征编码。这显然限制了图像特征的研究。特征表示学习 [14] 直接面向图像特征编码的研究，其指出从图像中提取的特征应当是各种语义属性的集合，不应当限制在某类特定的图像任务中。并且，表示学习提出良好的特征编码应当具有稀疏性，解耦性和可解释性。但在实际情况中，网络模型提取的特征编码往往难以满足上述的要求。客观评价所学习的图像特征的优劣是困难的，往往结合待解决的问题进行评估，即图像特征的编码依赖于特征的解码。这在端到端的深度学习模型中普遍存在。同时，图像特征编码依赖于学习方法的研究。CLIP 模型利用对比学习从图像和文本的成对数据中学习与文本特征对齐的图像特征编码。利用文本和图像之间的非精确标注进行学习的 CLIP 方法具有巨大的研究潜力。目前，许多研究工作都基于 CLIP 方法，利用其所预训练的图像特征编码来进行下游任务的处理。

图像特征解码：图像特征解码是利用图像特征编码解决指定任务所必要的环节。不同图像子任务间存在解码方式上的显著差异。例如，在图像分类任务中，图像特征解码可以看作为对图像特征进行分类的的判别器模型。在图像分割中，图像特征解码是利用获取的图像特征对原始图像的像素按类别进行划分的过程。在图像生成中，图像特征解码是基于图像特征到生成真实感伪造图像的映射。不同的图像任务往往对应着不同的解码器。为了研究特征在解码网络中的语义解码过程，本文所研究的图像特征解码，主要围绕基于特征编码的图像生成展开。深度学习中大致有三类生成模型，被用于处理图像生成问题。早期的生成模型的研究缺少与图像特征之间直接联系，其生成的过程是通过高斯隐变量分布来拟合真实图像的数据分布。自动编码器采用与特征编码器对称的网络结构来构建解码器，实现图像到特征编码，特征编码到图像的双向映射。由于缺少采样图像特征生成图像的过程，自动编码器无法直接利用特征编码来生成任意伪造的图像。变分自动编码器（VAE）[15] 参数化了图像特征空间，使其服从独立同分布的高斯分布，实现了特征编码的相互解耦，并可以在该空间中显示地采样生成伪造图像。然而，基于自动编码器的图像生成的研究难以

生成高分辨率的伪造图像。生成对抗网络解释了限制变分自动编码器的生成高质量图像的原因在于学习方式。变分自编码的自监督的学习方法无法指导编码模型生成高细节的图像。采用类似自动编码器的网络架构，GAN 引入了二分类判别器来分辨伪造图像与真实图像，用于解码网络的训练，从而提高了生成图像的质量。GAN 方法训练方法不够直观且稳定，难以高效地进行学习。最近，去噪扩散模型（Denoising Diffusion Probabilistic Models，DDPM）[16] 被提出，并作为新的生成模型收到了广泛关注。去噪扩散模型假设通过马尔可夫过程将随机高斯噪声不断添加到原图像样本上，使得图像最终服从标准的 $\mathcal{N}(0,1)$ 高斯分布。去噪扩散模型是建模在时间序列上的加噪过程。利用去噪扩散模型对输入图像的噪声进行预测，并按反向的时间序列迭代地对输入图像进行去噪，实现高清图像的还原。

然而，上述的生成模型都关注于图像生成的质量，对原始的隐特编码空间的假设都较为简单。这种简单的隐变量设置是难以直接应用于基于特征编码的图像生成问题。本文需要对原始的隐编码空间加以限制将其转换为条件生成模型，实现基于指定特征的图像生成。条件生成模型被普遍应用于一大类图像翻译问题上，如风格迁移 [17]、图像修复和图像超分辨率重建 [18] 等。这些研究都可以看作为从原图像提取图像特征，再通过解码网络对特征编码进行解码生成，并应用于具体的图像翻译任务中。

图像特征对齐：为了统一图像的特征编码和生成解码过程，需要将输入图像和生成图像进行数据分布的对齐。对于人脸图像而言，在解码生成模型中往往生成的是标准姿态的人脸，因此，在编码图像的过程中，对于任意输入的人脸图像需要对其进行必要的预处理，如裁剪和对齐等。然而，对于人脸图像受到拍摄角度和表情变化的影响，在二维图像中通过特征点定位的方法难以解决大角度偏差的人脸图像对齐。因此，我们试图研究了基于三维局部点云的对齐问题，并用于解决二维人脸图像的对齐。

为了解决成对三维点云的对齐，迭代最近点方法 (Iterative Closest Point，ICP) 等优化方法直接预测旋转矩阵和平移向量。然而，ICP 等诸多优化方法主要有两个方面的不足。其一，无法解决局部点的匹配。由于获取的角度存在偏差，成对的局部点云数据的匹配点获取是困难的。其二，无法解决全局的对齐。由于优化过程中会产生诸多局部极小值点，这导致难以获取全局的最优对齐结果。近年来，随着 3D 深度学习技术的发展，基于神经网络的方法被提出用以解决优化方法的缺陷。DCP [19] 实现了端到端的局部点云对齐，但只

在 ModelNet 数据集 [20] 上表现不错。DeepGMR [21] 采用混合概率分布进行点匹配，其参数由神经网络估计，并在 ICL-NUIM[22] 得到了验证。Deep Global Registration (DGR)[23] 提出了一个可区分的成对注册框架，包括特征对应预测、姿势估计和姿势细化。基于深度学习的对齐方法通常比迭代全局对齐方法运行速度快一个数量级，并且提高了全局对齐精度。然而，基于数据驱动的对齐方法往往只能在单一数据环境中表现较好，适应范围有限。深度学习对齐方法是对传统迭代优化对齐方法造成了巨大冲击。具有竞争力的迭代对齐方法是亟需新的研究成果。

图像特征编辑：图像特征编辑是理解图像空间与特征编码空间的核心任务，旨在通过编辑特征编码实现图像中特定属性的生成。基于图像特征的解码能够解决特征编码到图像的生成问题。该过程是隐式地通过神经网络模型完成的，缺少对图像特征与语义内容之间的探究及解释。在研究图像特征编码和解码的基础上，图像特征编辑需要深度分析特征编码空间和语义内容空间上的关联。首先，可视化激活特征层的研究表明提取的特征编码具有语义属性 [24]。在图像特征编码中，部分编码通道具有图像内容或属性相关的可解释性。例如，当特定目标在图像中显示时，对应的图像特征通道对此做出响应。图像特征未激活代表该特定的目标或属性在原图像中不存在。特征可视化的研究展示了图像特征与图像内容的对应关系。其次，基于图像特征的图像生成也具有类似的属性。不同图像特征通道能够生成不同的图像内容。研究者发现图像特征空间中的沿特定方向修改的原特征编码能够实现图像内容属性的定性修改，即通过修改特征编码可以对生成图像上的语义内容进行编辑。图像特征编辑的关键在于寻找图像特征与语义内容的对应关系，并确定编辑的特征方向以实现细粒度地控制生成内容。

人脸图像编辑研究存在许多挑战。第一，图像特征编辑方法难以直接评估。由于图像特征的编码和解码都面向指定的任务，研究的重心往往是任务性能和指标的提升。深度学习模型往往被视为黑盒模型，内部特征编码可解释性较弱，难以有效地评估。第二，图像特征编码是高维数据，难以发掘图像内容与相关图像特征通道的对应关系。第三，沿特定图像语义编辑，不仅需要理解图像特征的语义，还需要定位指定语义的特征编码通道，并寻找合适的编辑方法。针对以上问题，基于神经网络的图像编辑的研究主要分为两大类方法。第一类方法，通过控制网络模型的输入，来实现特定图像内容的编辑。该类方法大都采用条件生成对抗网络模型，并需要构造属性分类数据集，以此训练图像

编辑模型。例如，Wang 等人 [25] 训练了 Style 编码属性编辑网络，将原 Style 编码和指定的属性进行隐式融合生成新的 Style 编码，并利用 StyleGAN 中预训练的合成网络完成图像生成。EditGAN[26] 利用条件生成模型，用户可自定义输入图像的语义掩码，实现可人机交互的控制图像中不同物体的生成和修改。这些方法都通过神经网络隐式地进行属性内容的编辑，但对于更细粒度的可控编辑难以实现。第二类方法，试图发掘预训练生成模型的内部隐空间，寻找控制不同属性的特征编码，并对其进行编辑完成图像内容的修改。在人脸编辑问题中，InterfaceGAN[27] 通过线性支持向量机 (SVM) 在图像编码隐空间上找到关于人脸不同属性的分割超平面，并以此实现相关属性的编辑。StyleRig[28] 利用生成图像与原始图像的差异，在特征编码空间上找寻不同姿态、光照和表情等属性的编辑方向。

1.2.3 人脸编辑涉及学习方法的现状及难点

从上述关于图像特征的研究中可以看出，神经网络可被用于解决各种围绕图像特征的问题。对于不同的图像任务，网络模型结构和功能也存在显著差异。这导致了学习过程的困难，高效地训练神经网络模型成为了亟需解决的问题。随着越来越复杂的网络结构，新的学习机制被提出用于模型的训练。各种有效的学习方法被提出用于图像特征的相关研究。因此，本文回顾了其中主要涉及到的学习方法，大致包含有自监督学习、生成对抗学习、对比学习和知识蒸馏学习等。

自监督学习：自监督学习 [29] 主要利用辅助任务（pretext）从大规模的无标注数据中挖掘自身的标注信息，通过构造的监督信息对网络进行训练，从而学习到对下游任务有价值的特征编码。在图像的相关任务中，学习图像的高级语义特征对下游任务是十分重要的。迁移学习的研究表明将相关任务的模型进行迁移，在其他任务上可以有效减少对标注数据的需求。然而，这些神经网络模型学习到的图像特征是基于标注数据的监督学习方法，无法利用未标注数据。自监督学习试图从无标注的数据构造监督信息，从而训练模型学习高级的语义特征。按构造数据的标注信息区分，自监督学习的方法大致可以分为两类：1）基于上下文信息的方法，2）基于数据增强的方法。基于数据本身的上下文信息，在自然语言处理的词向量预训练模型上展现了强大的学习潜力。在词向量的预训练中，连续词袋模型（CBOW）通过句子前后的词来预测中间的词向量，而 Skip-Gram 方法通过中间的词来预测前后的词向量。BERT 方法 [30]

将句子中的单词进行随机遮盖处理，并学习词向量用于预测掩码。借鉴 NLP 自监督预训练的方式，研究人员试图利用无标注图像学习特征编码。Jigsaw 拼图方法 [31] 将原始图片切片成合适大小的图像块，利用网络学习图像特征完成对图像块的按序拼接。Noroozi 等人 [32] 通过预测输入成对图像块间相对的位置关系，学习图像块的特征编码。论文 [33] 利用视频提取的帧图像，将相邻帧作为正样本，并引入噪声负样本。通过缩小正样本和增加负样本间的编码距离，训练网络学习图像特征。由不同视角下获取的相同目标的图片集也可以作为图像在空间中的上下文。Tsai 等人 [34] 提出了多视角下的图像特征的自监督学习方法。与基于上下文信息的方法不同，基于数据增强的方法通过对图像进行变换来生成成对的标注信息，实现相似图像的特征学习。例如，利用黑白与彩色图像间的转换，张等人 [35] 学习了图像中具有相似结构的特征编码。通过转旋图片的方法，Gidaris 等人 [36] 预测原始图像到目标图像间的旋转角度，实现了语义空间上的特征对齐。然而，从自监督学习方法研究目标可以看出，利用无标注的数据构造的监督信息受限于其数据本身，设计更复杂的关于语义处理的辅助任务仍旧是困难的。这限制了自监督学习方法对图像特征的学习能力。

生成对抗学习：生成对抗学习是一种非监督的，通过让两个或多个神经网络模型相互博弈的方式进行学习的方法。生成对抗网络是应用生成对抗学习方法进行数据生成的网络模型。生成对抗网络通常由一个生成网络与一个判别网络组成。从独立同分布空间的高斯模型中随机取样作为生成网络的输入，其输出结果需要尽量模仿训练集中的真实样本。判别网络则判断输入样本来自于真实样本或生成网络的伪造样本，并将伪造样本从真实样本中尽可能分辨出来。生成网络则要生成尽可能逼真的样本欺骗判别网络。两个网络的目标相互对抗，训练过程不断调整其参数实现各自任务的学习。由于可以直接利用未标注的图像，生成对抗网络被广泛应用于图像生成任务中。图像翻译 [37] 是其中典型的图像生成任务，且较为适合研究图像特征的编解码过程。图像翻译的目标是将图像中的内容从一个图像域转换到另一个图像目标域。在图像翻译任务中，通常采用条件 GAN 模型实现源图像域到目标域的迁移。图像翻译为确保图像域之间转换内容的一致性，图像内容与特征编码要实现对齐。例如，语义分割需要实现图像到语义掩码的边缘轮廓的对齐，而相似物体图像翻译只需目标物体的特征编码对齐，并保留其余部分的内容等。在图像生成的方法中，生成对抗网络对比自监督学习下的 Autoencoder 生成模型能生成更真实的图像。ProGAN[38] 提出了从低分辨率到高分辨率的学习过程，实现了从 256×256 到

1024×1024 分辨率的图像生成。借鉴了风格迁移内的 Style 风格，StyleGAN[39] 将独立同分布的高斯随机变量 z 嵌入到隐式的 Style 空间，并利用 Style 特征编码实现逐层上采样的图像生成。GANspace[40] 和 StyleFlow[41] 的工作表明对比原始的高斯分布空间，隐式的特征编码空间具有更好的特征解耦性质，这为后续的基于图像特征编辑技术实现可控的内容的图像生成做出了重要铺垫。

对比学习：对比学习 [42] 着重于学习同类样本之间的共同特征，区分非同类样本之间的差异特征。其中，对于指定中心点样本，如何选取合适的正样本和负样本是面临的主要问题。对于序列数据，CPC[43] 将相邻序列片段上的作为正样本，随机采样多个其他序列的数据作为负样本，并提出了 InfoNCE 损失用于训练网络。与序列数据不同，图像数据需要从数据集中选取正负样本。为实现从数据集中选取正负图像样本，可以利用图像特征进行匹配选取。为加速高维图像特征地匹配，MemoryBank[44] 动态记录了大量图像的高维语义特征，以此为训练样本提供正负样本。MoCo[45] 提出了动量更新 MemoryBank 中图像的语义特征编码，解决 MemoryBank 特征编码难以及时更新的问题。SimCLR[46] 利用大规模地分布式训练集群，实现了大批量的图像特征编码匹配。在每个训练批次中，能够实现对所有样本对的图像特征的距离计算，加速了训练过程。比较生成对抗学习，对比学习不需要关注生成实例上繁琐的细节，只需要在语义特征编码上对数据进行区分。对比学习更关注于图像特征本身，学习到的特征编码具有更强的泛化能力。同时，自监督学习与对比学习也存在着交集。自监督学习中可以通过对原始数据进行增强生成对应的正负样本对，并利用对比损失函数进行学习。SimCLR 通过简单的数据增加方法实现了对比学习，可以看作为这两种学习方法的结合。DIM[47] 引入了对比学习来训练辅助图像任务模型，以实现无标注数据上的自监督学习。Tian 等人 [48] 提出了利用多模态学习生成正负样本对，训练编码网络学习图像的特征。

知识蒸馏学习：知识蒸馏学习 [49] 侧重于学习高性能的小规模网络模型，旨在将学习能力强的复杂教师模型中的“知识”迁移到简单的学生模型中。狭义上的知识蒸馏学习方法，教师模型和学生模型所面向的待解决问题是一致的。然而，在模型间方法学习的研究中，知识蒸馏学习方法可以利用教师模型的知识并用于指导学生模型的学习。从模型间的学习角度出发，广义上的知识蒸馏学习方法是一种训练学生网络模型的有效手段。其中，教师模型和学生模型可以分别面向不同的任务。对于难以利用标注数据进行学习的任务而言，教师模型可用于指导学生模型的训练。这在人脸编辑和风格迁移等难以了量化生成结

果的问题上，知识蒸馏学习方法被应用于众多提出的解决方法中。对比生成对抗学习方法和对比学习方法，知识蒸馏学习方法也是面向模型间的学习过程。知识蒸馏学习方法面向学生模型和教师模型间的指导学习，生成对抗方法面向学生模型间的合作学习。

从上述的介绍可以看出不同的学习方法所研究的侧重点不同，彼此之间没有清晰的区分界限。对比学习、自监督学习和生成对抗学习之间有着紧密的联系，彼此相互借鉴和发展。围绕图像特征相关的人脸编辑方法可以被同时分类到上述的学习方法中。

1.3 本文工作

现实世界中采集到的人脸往往不能直接用于人脸编辑，我们提出了一个通用的人脸编辑框架，用以系统化处理人脸编辑过程。结合1.2中人脸编辑面临的问题和挑战，我们将人脸编辑框架总结为图1-1中的四个主要流程。如图1-1所示，本文将从四个方面展开研究，分别围绕图像对齐、图像特征编码、图像特征解码和图像特征编辑展开。人脸图像对齐、图像特征编码、解码和编辑采取递进式的研究思路，逐步深入探究图像特征，并利用其解决人脸图像的属性编辑问题。具体的研究内容包括：

1. 图像对齐旨在统一图像的编解码过程，是后续在特征编码上进行编辑的预处理步骤。由于图像编解码被分别独立地研究，利用局部点云的对齐任务，端到端地实现输入图像与生成图像中语义内容的对齐。人脸编辑主要针对标准姿态的人脸进行处理。对于任意获取的人脸图像，将其对齐到标准姿态的预处理步骤，有利于上述人脸图像的编解码过程。我们研究了成对局部点云的对齐及补全任务，并提出了基于混合优化的点云对齐方法。该方法利用无约束变量在整个变换矩阵空间进行迭代优化，解决了点云的局部和全局匹配，提高了对齐精度。该研究可以辅助进行人脸图像对齐，为后续人脸特征编码、解码生成和编辑进行铺垫。
2. 图像特征编码是神经网络模型解决图像编辑的核心研究内容。目前，学习具有强泛化能力的图像编码依赖于海量的标注数据。然而，标注海量数据是十分困难，甚至在某些图像任务中是不可行的。立足于无标注的图像数据，本文结合图像聚类任务致力于无监督地学习图像特征编码。我们提出了 DERC 聚类方法，能够有效地学习图像特征编码。该研究初步研究了人

脸编辑问题中图像的特征编码。

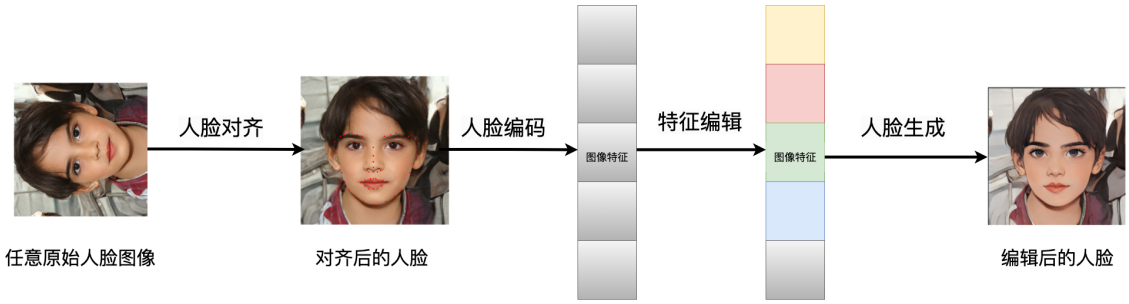
3. 图像特征解码是解决图像编辑问题中的关键环节。本文主要关注特征解码中基于图像特征的生成任务。本文分析了预训练的 StyleGAN 图像生成模型中 Style 特征编码的解码过程。基于 Style 编码的生成模型加强了编码空间到图像空间的可解释性，提高了生成图像的质量。该研究初步分析了基于 Style 编码的合成网络在人脸生成时的解码过程。我们后续提出的人脸编辑技术便是受此解码过程的研究启发的。
4. 在图像编解码研究的基础上，本文探索了图像编码和语义内容之间的映射关联，在编码隐空间上实现了对指定语义内容的编辑。针对人脸属性任务，在上述 StyleGAN 中隐编码空间 $\boldsymbol{S}$ 定位与指定属性的通道。最后，还提出对特征编码的单通道和多通道编辑方法，实现细粒度地控制生成内容。该研究利用合成网络的分层 Style 编码提出了可指定人脸属性的编辑方法。

如图 1-1所示，上述的研究内容及解决方法分别被用于脸图像编辑的各个阶段中。首先，局部点云对齐的研究用于人脸图像对齐，作为预处理中关键步骤。目前，神经网络应用于人脸图像编辑，主要围绕标准姿态的人脸进行展开。我们提出了高精度的局部点云对齐方法，通过提取二维人脸图像中的关键特征，实现人脸标定及对齐。其次，图像特征编码的研究用于人脸图像编码，实现人脸到特征编码的映射。本文结合聚类任务，提出了无标注数据的图像特征编码。这有利于实现任意人脸图像的编辑。同时，图像特征解码的研究用于分析人特征到人脸图像的生成过程。针对预训练的人脸生成模型，探索了特征编码与人脸属性的关联关系，明确了特征编码空间和人脸图像语义空间的对应关系。最后，通过在人脸特征空间的分析，利用特征编码完成可指定人脸属性内容的编辑。本文提出了人脸特征编辑方法，通过结合人脸图像特征编解码过程，可实现多种人脸属性编辑。

1.4 论文结构

本文主要围绕深度学习下的人脸图像编辑问题进行研究，主要的内容包含了人脸对齐、图像特征编码、图像特征解码以及图像特征编辑等四部分研究内容。全文共分为七章，每一章的具体内容如下：

第一章，绪论，介绍了深度学习下人脸图像编辑的研究背景和研究意义，并分析了当前基于图像特征的人脸编辑的研究现状。



人脸编辑流程示意图

图 1-1: 人脸编辑流程示意图。

第二章，预备知识与相关工作。主要介绍了图像特征研究中所采用的相关的神经网络模型及其结构。

第三章，人脸编辑下的特征对齐研究。借助三维的点云数据，研究不同视角下的局部点云对齐问题。本文提出了高精度的局部点云的对齐方法。该研究能够被用于解决人脸图像的对齐问题。

第四章，人脸编辑下的特征编码研究。解决了无标注图像数据下训练网络编码器学习图像特征，初步解决了人脸图像的编码问题。

第五章，人脸编辑下的特征解码研究。研究基于特征编码的图像生成过程。基于预训练的 StyleGAN 模型，初步解决了基于 Style 编码的人脸生成问题。

第六章，可指定属性的人脸编辑研究。基于图像特征编码解码和人脸对齐方面的研究，提出通过特征编码实现对人脸属性进行编辑的方法。

第七章，总结并展望了围绕人脸图像编辑的研究内容。

第二章　相关工作

2.1　引言

在图像处理问题上，深度神经网络能够端到端地提取图像特征，并对其进行有效地解码，解决各种子任务。各种不同结构的网络模型被提出，用以学习不同层级含有语义信息的图像特征。神经网络的研究者致力于提出新的网络模块层，用于改进基础的网络架构，提升网络对图像特征的表示能力。借鉴传统的特征提取方法，卷积常被用来在空间域和频率域中提取图像特征，并被成功应用于神经网络模型中。得益于卷积操作的研究，卷积神经网络成为主流地处理图像问题的模型，并具有着良好的性能。图像任务上的模型大都基于卷积神经网络搭建，并在此基础上针对具体子任务对网络进行必要改进。例如，图像分类作为图像的基础任务，许多研究致力于提出新的计算模块和结构，用来改进原有卷积神经网络的不足，如 Normalization 机制 [50]，Attention 机制 [51] 和跳层连接等。这些优秀的改进模块不仅提升模型在分类任务上的性能，还具有一定的泛化性可被用于其他任务的模型搭建中。因此，本章首先介绍了主要的神经网络类型，其被广泛应用于各种图像任务中。其次，介绍了常用于处理图像任务的网络结构，其主要包括自编码器网络和生成对抗网络两种模型结构。最后，围绕图像特征的不同研究内容，详细介绍了相关的图像研究任务和已有的解决方法。

2.2　网络类型

深度学习模型，即深度神经网络模型，通常由多层的网络层堆叠而成。按照每层神经计算单元的结构划分，可以分为全连接神经网络、卷积神经网络、循环神经网络和 Tranformer 神经网络等。在图像处理中，由于传统方法普遍采用卷积来提取图像特征，受此启发的卷积神经网络成为处理图像任务的主流架构。卷积神经网络展示了可靠的图像特征学习能力。最近，基于 Transformer 的神经网络在部分图像任务上超过了卷积神经网络，成为了新的有待发掘的模

型。后续本文的图像特征研究中，许多方法的模型都采用了上述的两种类型。因此，本节对这两类网络模型进行相关的介绍。

2.2.1 卷积神经网络

卷积神经网络（Convolution Neural Network，CNN）是指包含有卷积层的前馈神经网络。通过堆叠卷积层可以搭建简单的深度卷积神经网络。卷积操作就是利用卷积核（卷积模板）将图像上的像素灰度值与选取的卷积核进行矩阵点乘，然后将所有相乘后的值相加作为卷积核中间像素点计算的输出灰度值。相较图像的尺寸，卷积核尺寸较小，卷积操作需要沿着图像像素进行滑动，完成对所有像素点的计算过程。在图像处理上，卷积核通过滑动参数共享的形式是十分合理且有效率的。卷积层不仅减少了模型的参数量，也解决图像中物体的平移不变性问题。同时，在传统的图像处理上，卷积核也被称为滤波器。不同的卷积核对应着不同的功能，具有良好的可解释性。例如，高斯卷积核实现了图像的平滑。Sobel 卷积核能提取图像的边缘。拉普拉斯卷积核能实现图像的锐化。卷积神经网络则通过训练的方式，自动学习不同类型的卷积核参数。通过卷积层和非线性激活层的堆叠，能够提取更复杂关于图像内容的语义特征。

卷积操作：卷积作为一种数学运算，被广泛应用在信号分析中。在连续的一维函数上，设 $f(x), g(x)$ 是 $\mathcal{R}$ 上的两个可积函数。两个函数的卷积可以由 $g(x)$ 关于原点翻转并滑过 $f(x)$ 函数，并计算每个位置的积分。卷积算子由 $*$ 表示，具体的一维连续函数的卷积定义如下：

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(\tau) g(x-\tau) \mathrm{d}\tau. \tag{2-1}$$

其中，x 代表滑动的位移，τ 是积分假变量。卷积由许多重要的性质。卷积定理揭示了空间域的卷积和频率域中乘积的变换关系。设 $F(\mu)$ 和 $H(\mu)$ 为 $f(x)$ 和 $h(x)$ 的傅立叶变换，卷积定理如下所示：

$$\begin{aligned} f(x) * h(x) &\Leftrightarrow F(\mu) H(\mu), \\ f(x) h(x) &\Leftrightarrow F(\mu) * H(\mu). \end{aligned} \tag{2-2}$$

其中双箭头指示左边的表达式通过右边的表示式的傅立叶反变换得到，右边的表达式通过左边的表达式的傅立叶变换得到。卷积定理表明，空间域中两个函

数卷积的傅立叶变换等于这两个函数傅立叶变换在频率域的乘积。空间域中的两个函数乘积的傅立叶变换等于这两个函数的傅立叶变换在频率域的卷积。卷积操作能有效提取 $f(x)$ 关于 $h(x)$ 的特征。

在处理图像问题中，图像由二维平面上的离散的像素点组成。将式 (2-1) 中的卷积拓展到二维空间上的的离散变量，其卷积操作定义如下：

$$f(x,y) * h(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n)h(x-m, y-n), \tag{2-3}$$

其中 $x \in [0, 1, \cdots, M-1], y \in [0, 1, \cdots, N-1]$。在式 (2-3) 中，$f(x,y)$ 可以代表图像 I，$h(x,y)$ 代表卷积核。然而，卷积神经网络中，空间域的卷积核通常只处理图像中的局部信息，所构建的卷积核大小为 5×5，3×3 甚至为 1×1 等。模型利用多层的卷积层串联实现扩大感受野，来提取图像的局部和全局特征。因此，在图像卷积的处理中，需要对每个像素点区域进行二维平面上的卷积，其具体流程如下：

$$I * g(i,j) = \sum_{u=-k/2}^{k/2} \sum_{v=-k/2}^{k/2} I(i+u, j+v)g(u+k/2, v+k/2) \tag{2-4}$$

其中 I 代表图像，g 表示大小为 k 卷积核。图像上的卷积操作在以每个像素位置 (i, j) 的中心的 $k \times k$ 区域上进行离散的二维平面上的卷积运算。由于式 (2-4) 中卷积核并没有做翻转操作，其实并不是严格意义上的卷积操作，更确切的描述应该是指空间相关的滤波操作 [52]。

卷积网络模型：大小为 $H \times W \times 3$ 的彩色 RGB 图像经过卷积计算后输出为 $H/s \times W/s \times C$ 的特征图，其中 s 为卷积核在图像上的平移步长。卷积神经网络内部中每层都对特征图进行卷积处理提取特征，并输出对应的特征图。深度卷积神经网络可以看作特征图在各个层中的信息处理和传递。设 a_i 为第 i 层网络的特征图输入，则 a_{i+1} 为第 i 层的输出。由卷积层搭建的 n 层神经网络模型，其具体形式化定义如下：

$$a_{i+1} = \delta(Conv_i(a_i)), \forall i = \{0, 1, ..., n-1\}. \tag{2-5}$$

其中 $Conv_i(\cdot)$ 代表着第 i 层的卷积操作，$\delta(\cdot)$ 为非线性激活函数。由于卷积操作由矩阵乘积表达，其本质属于线性运算。非线性激活函数的引入是为了利用多层卷积网络实现对非线性函数的拟合。理论上，含有两层的具有非线性激活函

数的多层神经网络可以拟合任意连续函数。三层的神经网络可以构建非凸或不连续的决策边界 [53]。然而，理论上的研究在现实中往往是难以达到的，浅层网络模型难以直接表示复杂函数。在相同计算神经元下，深层网络比浅层网络具有更丰富的表达能力。在处理图像问题中，模型的卷积操作受限于卷积核大小，只能提取图像的局部特征。为捕获大范围的图像特征，模型需要构建多层卷积层来扩大感受野。通过堆叠多层的卷积层，使得模型能够表达丰富的图像特征。然而，深层网络和海量参数使得训练模型成为难点。如何有效从数据中学习网络参数成为了亟需解决的问题。

卷积神经网络在图像识别等方面取得了巨大的成功。VGG 网络 [10] 构建并学习了含有 16 或 19 层的深度神经网络模型，验证了增加网络的深度能够在一定程度上提高模型性能。ResNet[1] 提出了跳层连接结构，将可学习深度扩展到了上百层，并加速学习深层神经网络的过程。同时，跳层连接对网络中提取的图像特征能有效地进行保留。EfficientNet[11] 通过神经网络架构搜索模型结构，实现多种卷积模块的优化组合，减少了对层数的依赖并提升了推理性能。虽然卷积神经网络在图像识别上不断刷新着性能指标。但对网络内部的图像特征如何提取，如何理解各层中的图像特征图的意义等问题的研究是滞后于模型发展的。为理解网络内部学习到的卷积核和提取的图像特征，Olah 等人 [24] 可视化展示了卷积神经网络的各层的特征图的语义内容，并对此进行了解释。网络内部是分层次获取图像特征的。按照不同层划分，每个层在提取图像特征中都有着相对粗糙的解释。模型逐层提取到了图像的颜色、线条、纹理和局部目标等语义特征。然而，关于图像特征和图像内容的更精确细致地解释有待进一步的研究。

2.2.2 Transformer 网络

Transformer 网络起源于自然语言处理，并被逐步应用于图像问题的处理。Transformer 网络不同于卷积神经网络和循环神经网络，其由堆叠的 Transformer 模块组成。在图像分类任务上，ViT[3] 成功利用 Transformer 网络模型超过了卷积网络模型的分类性能。这给予了采用 Transoformer 网络解决图像问题的新思路。

Transformer 模块：Transformer 单元是组成 Transformer 网络的核心模块，用以代替卷积操作来提取图像的语义特征。类似于自然语言中捕获上下文的语义信息，将图像按照像素块划分及序列编码后，Transformer 单元可以捕获图像

块间的全局语义信息。对比卷积神经网络需要通过多层的卷积操作来扩增感受野，Transformer 网络在全局信息的获取效率上优于卷积神经网络。Transformer 单元可以看作为全连接网络的 Attention 版本，每个单元仅由 Attention 模块和前向全连接网络组成。其中，Attention 机制是计算输入 Query 向量对输入 Key 向量的关注度，按照得分乘以输入 Value 向量，并依据 Query 向量输出查询结果。设 Q 为 Query 矩阵，K 为 Key 矩阵，V 为 Value 矩阵。Attention 的形式化表达为：

$$Attention(Q, K, V) = \text{Similarity}(\frac{QK^T}{\sqrt{d_k}})V. \tag{2-6}$$

其中函数 Similarity(.) 为相似度计算函数，d_k 与输入矩阵 Key 的维度相关。依据相似度计算函数的不同，可以划分不同类型的 Attention 实现方法，如 Soft-Attention 和 Hard-Attention 等。图像处理中常采用 Soft-Attention 方法。Soft-Attention 利用 Softmax 函数来计算 Query 向量对 Key 向量的得分，取值在 $[0, 1]$ 之间。Soft-attention 可以表示为

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{2-7}$$

其中 d_k 为 K 矩阵列向量的维度，用于归一化查询结果，使其符合标准方差分布。在 Transformer 模型的 Attention 计算中，Query 向量，Key 向量和 Value 向量都来自与图像块序列。图像块序列中每个位置与其他位置计算相似度，通过 *Attention* 模型函数加权得到全局编码特征 Z。编码特征 Z 是图像特征 Q 计算全局 Attention 后的中间结果。为了解释并分离 Attention 特征 Z，Transformer 单元通过两层全连接网络（Feed Forward Neural Network，FFN）将 Attention 提取的全局特征进行非线性变换加以分析，其可以表示为：

$$FFN(Z) = \max(0, ZW_1 + b_1)W_2 + b_2, \tag{2-8}$$

其中 (W_1, b_1) 和 (W_2, b_2) 分别代表第一二层全连接网络的参数。

Transformer 网络模型：与搭建卷积神经网络类似，通过堆叠 Transformer 模块可以搭建 Transformer 网络模型。设图像分块后的图像序列化特征为 a_0，大小为 $N \times K$。这里 N 代表图像块序列长度,K 为初始图像块的特征维度。Transformer 模块将输入特征 a_i 分别计算 (Q, K, V) 并进行全局 attention。利用全连接 FFN 模块经激活函数得到输出特征 a_{i+1} 并作为下层的输入。由 Transformer

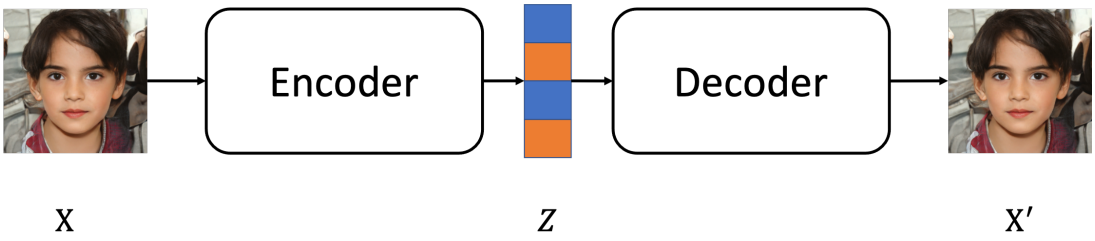


图 2-1: 自编码器网络示意图。

模块搭建的 n 层网络模型，其具体形式化定义为：

$$a_{i+1} = \delta(Trans(a_i)), \forall i = \{0, 1, ..., n-1\}. \tag{2-9}$$

其中 $Trans$ 函数代表 Transformer 模块计算单元，由上式 (2-7) 和 (2-8) 组成为:

$$Trans(a_i) = FFN(Attention(a_i, a_i, a_i)). \tag{2-10}$$

通过堆叠 Transformer 单元，研究者构建的 Transformer 网络可被用于众多图像任务。ViT[3] 只采用 Transformer 单元搭建的图像分类网络模型超过了同时期的卷积神经网络。Segmeter[54] 将 ViT 模型拓展到图像语义分割任务，实现对图像特征的内容语义的解码。MaskGit[55] 成功利用 Transformer 网络进行快速地图像生成。

2.3 网络结构

2.3.1 自编码器网络结构

自编码器网络（AutoEncoder，AE）由编码器（Encoder）和解码器（Decoder）组成。如图 2-1所示，编码器主要将输入 X 转换成中间语义特征 Z，然后解码器将 Z 转换成与输入相似的输出 $\overline{X}$。自编码器网络旨在通过学习使得输入 X 和输出 $\overline{X}$ 无限接近。当输入 X 的维度高于编码 Z 的维度时，编码器可以看作为对输入 X 的降维过程，实现高维数据到低维数据的映射。当自编码器用于图像问题时，编码器和解码器大都为卷积神经网络。通过自监督的学习方式，图像能够学习低维的数据表示 Z。低维图像特征 Z 可以看作为一种图像特征编码。不同于监督学习下的图像任务模型，自编码器网络利用自监督学习方法，能够从无标注的数据学习图像特征。在机器学习中，自编码器常被作为一

种无监督学习方法，并用于数据降维。然而，在深度学习中，自编码器不仅可以实现对图像的降维，也可以实现图像的特征提取。编码器可被用于从无标注数据中学习图像特征，可被用于研究图像特征的编码问题。解码器可以实现特征到图像生成的解码，可被用于处理基于特征的图像生成问题。自编码器常被看作为两个不同功能模型的组合。编码器用于解决入图像编码问题，解码器用于解决图像生成问题。这两个问题可以看作为互为对偶问题。我们还可以进一步研究特征与语义内容的潜在联系，处理基于特征编码的图像编辑任务。在之后的章节，本文会详细探讨自编码器网络的学习方法以及上述围绕图像特征的研究。

2.3.2 生成对抗网络结构

生成对抗网络（Generative Adversarial Networks，GANs）是由生成器（Generator）和判别器（Discriminator）组成。如图2-2所示，生成器用于图像生成。判别器用于判别伪造图像和真实图像。为了实现图像的采样和伪造，生成对抗网络从易于采用的先验高斯分布中生成变量 Z，生成器 $G(Z)$ 输出对应的生成图像 X。判别器为二分类模型推断 $D(X)$ 为真实图像的概率。原始的 GANs 模型先验分布是易于采样的，容易伪造丰富的逼真图像。然而，高斯分布难以直接与图像特征进行关联。由于本文主要围绕图像特征展开研究，随机从高斯分布采样的噪声数据 Z，显然无法将其作为一种有效的图像特征。因此，部分研究者将特征编码引入到图像生成模型中。一种简单的方法是将 GAN 模型与上述的自编码网络结合，将编码器提取的特征作为条件输入控制生成器的解码过程。通过条件的输入特征，条件生成对抗网络实现将图像编码特征融入到图像生成解码过程。条件生成对抗网络在基于特征编码的图像过程中被广泛应用。例如，在人脸变换和风格迁移任务中，大部分方法都利用条件生成对抗模型，实现原图像域到目标图像域的图像翻译。其中，图像特征在编解码过程中起到关键作用，确保生成内容的对应。

2.4 围绕人脸编辑的若干研究任务

本文主要关注深度学习下图像特征的研究，并围绕图像特征的编码、解码和编辑三个部分展开。由于深度学习下的图像特征研究大多与图像任务相关，本节将主要介绍与研究图像特征相关的图像问题。具体的内容如下，与图像特

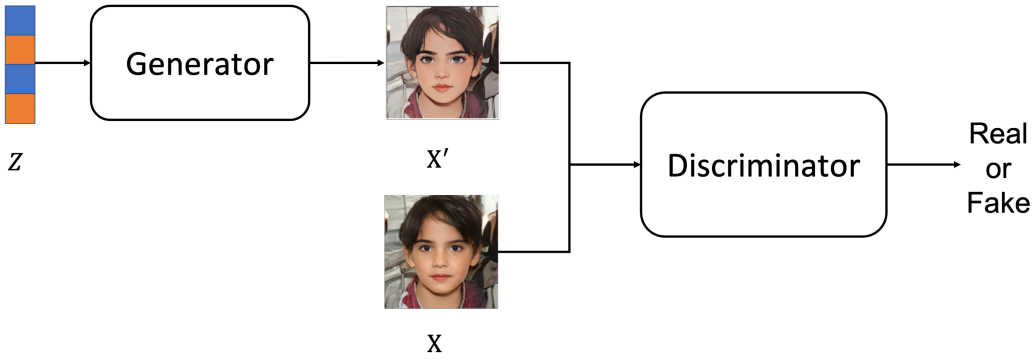


图 2-2: 生成对抗网络示意图。

征提取编码较为密切的图像分类问题，基于特征编码的图像生成问题和探究特征编码空间与语义内容的图像编辑问题。需要指出的是，图像特征的研究是可以独立于具体的图像任务的。因此，本文阐述相关图像任务时，尽可能的围绕图像特征展开，而忽略不同图像任务上的差异。

2.4.1 人脸特征编码学习任务

图像特征编码是利用深度模型从图像中获取有着特定语义内容的过程。当网络模型训练完成后，从图像中抽取的内容为特征表示，也被称为特征编码。即原始图像 X 被网络模型映射到编码空间中的 Z 编码。在图像分类任务中，全连接层之前的多层卷积网络模型常被作为通用的图像特征编码器，能有效提取具有良好区分性的高级语义特征，并被直接应用于其他图像问题。然而，这种依靠图像分类任务学习的特征提取器有两个主要问题。第一，所获取的图像特征编码是面向图像分类任务而言的。这种编码的方式丢失了部分表述原图像的能力，难以胜任不同的图像任务。例如，由图像分类学习的特征编码无法适用于在解码过程中的图像生成问题。第二，图像分类任务的学习过程依赖大量的手工标注数据。基于 Transformer 的 ViT 研究表明通过增加标注的样本来训练模型，可以学习到更具有泛化性能的图像特征 [10]。标注的图像数量制约了其特征编码的能力。

针对图像分类任务中的问题一，迁移学习提出了一种有效的解决方案。不同的图像任务可以利用预训练加微调的方式来优化原特征提取器的图像编码。finetune 方法 [56] 有效解决了不同任务间所关注的图像特征的差异。同时，迁移学习加速了模型学习收敛的过程。图像分类任务的预训练特征编码器也有效地避免了因目标任务数据不足带来的过拟合的风险。finetune 方法和迁移学习

等都是面向图像任务的，并主要作用于图像特征编码。在相关任务中，finetune 方法能明显改善所提取的图像特征，实现高效地语义编码。为了解决在问题二中手工标注数据的困难，CLIP[13] 借鉴了自然语言处理中的预训练词向量的方法。在收集的图像文本对的匹配任务中学习图像的特征编码。在该任务中，CLIP 利用大批量的训练数据来构建文本和图像的匹配目标，实现了文本编码和图像编码器的自动对齐。由于图像文本对都采集自互联网数据无须人为标注，CLIP 解决了无标注图像中特征编码的学习。模型训练过程中利用对比学习方法有效解决了图像分类任务中的问题二，这拓展了图像特征编码学习新的研究任务。同时，StyleCLIP[57] 等研究表明 CLIP 的图像特征编码具有着丰富的可解释性。图像生成任务中的解码过程能够直接基于 CLIP 的图像特征编码展开后续的研究。

2.4.2 基于语义特征的人脸图像生成任务

基于图像特征的图像生成任务是图像特征编码的逆过程，其试图通过图像特征的编码将其映射回原图像空间。结合特征编码的内容，该图像特征解码的研究实现了原图像到特征编码，特征编码到原图像的闭环。因此，本文所研究的图像特征解码过程是基于图像编码展开的。基于语义特征的图像生成任务常采用自编码器网络结构。通过中间隐藏的图像特征编码，利用重构后的图像与原图像进行自监督学习。然而，利用自编码器的网络结构生成的图像较为模糊，难以与真实图像混淆。在图像生成任务的研究上，首先研究的关注点在图像生成的质量和多样性上。自编码器网络生成图像质量较差且难以直接采样。针对上述不足，变分自编码显示地学习图像特征编码使其服从高斯分布，简化了采样过程。为进一步提高图像的生成质量，研究人员引入 GAN 网络利用生成器伪造图像，并通过判别器进行生成对抗学习。最近，扩散模型（Diffusion Mode）[16] 通过对生成函数的梯度建模，利用梯度下降的方法来迭代生成逼真的图像。这种方法虽然易于训练且生成的图像质量上堪比 GAN 方法，但其在图像生成效率上差一至两个数量级。然而，上述的这些方法对图像生成过程中的输入的图像特征与生成内容的联系关注较少，缺少对语义特征在解码时的分析。

与基于图像特征的图像生成较为密切的研究是一类图像翻译任务 [58]。图像翻译任务是实现源图像域到目标域的转换。许多计算机视觉问题都可以被看作图像到图像的转换任务。例如，CycleGAN[59] 提出了通用的图像翻译模型，

并被用于标签图到场景图的转换、线条轮廓到色彩图像转换、图像的风格转换，春夏场景变换等各种图像应用场景。同时，语义分割 [60] 和风格迁移 [17] 作为两类特殊的图像翻译任务被单独地加以研究，是图像翻译发展过程中两类重要的应用领域。语义分割将标签或类别与图片的每个像素进行关联，用来划分构成可区分类别的像素集合。语义分割可以被视为将图片映射到语义掩码的图像翻译任务。风格迁移的目标是对内容图像用将新的风格重新进行表达，生成新的混合图像，例如风景照到油画之间的转换，人物头像到动漫头像的风格迁移。众多的图像翻译子任务都可以看作为是条件生成模型的不同问题上的具体应用。利用条件生成模型可以方便研究图像特征的解码过程。

2.4.3 人脸图像特征编辑任务

图像特征编辑是理解图像特征空间与语义内容的关键任务，旨在修改图像特征来指导特定图像属性的生成。图像特征编辑任务需要实现对生成图像的细粒度控制，研究图像特征的解耦和解释，并完成对图像的编辑。在图像编辑任务中，由于人脸生成及编辑 [61] 具有重要的应用和娱乐价值，对其相关地研究受到了广泛关注。在人脸生成任务中，不同于以往在高斯噪声分布上直接伪造人脸图像，StyleGAN 将随机噪声映射到风格编码隐空间，提高了生成图像的稳定性和质量。为了在隐空间实现对图像特征编辑，InterFaceGAN[27] 和 StyleSpace[62] 等方法研究并分析了其各种潜在的特征编码空间，如 W 空间和 Style 空间，并发现对该空间进行特定编辑与生成图像的语义内容相关。InterFaceGAN 揭示了不同性别和年龄的人脸在 W 特征编码上具有特定的编辑方向与之对应，修改关联的特征编辑方向可以控制人脸属性的生成。StyleSpace 在 Style 编码空间实现了多种人脸的细粒度编辑，例如头发，眼睛，嘴巴等内容。Encoder4Editing[63] 表明可以直接利用编码网络，将图像映射到对应的 Style 空间以实现原图像到目标图像的翻译。类似 Encoder4Editing 工作，Pixel2Style2pixel[64] 的图像编码重建研究也展示了每张图像通过编码器将其映射到 Style 编码，并通过预训练的 StyleGAN 模型进行恢复重建。因此，Style 编码可以被视作为一种图像特征编码。本文将关注 StyleGAN 生成模型，并研究各种潜在的图像特征编码对图像语义内容的联系，提出特征编辑方法细粒度地控制图像的生成内容。

2.5 本章小结

本章回顾了深度学习中处理图像的基本网络模型，分别为卷积神经网络和Transformer 神经网络。目前，大多数关于图像网络模型的研究大都围绕这两种神经网络展开。对比由 Transformer 模块搭建的网络，卷积神经网络在特征编码上具有更好的可解释性。因此，本文主要围绕卷积神经网络上的图像特征展开后续研究。其次，介绍了不同的网络结构。随着图像问题越来越复杂，单一的网络模型难以学习和处理待解决的任务。研究人员将多个不同功能的网络进行组合，构建更有效地学习模型。我们讨论了两类网络结构，分别为自编码器网络结构和生成对抗网络结构。这些结构中都将整个模型按照基本功能进行了网络划分。通过对上述模型的结构分析，有助于理解本文后续研究图像特征中的相关模型。最后，本文讨论了图像特征在不同图像任务下的侧重点，分别为图像语义编码，基于图像特征的语义生成和图像编辑等相关任务及研究内容。图像特征编码和生成解码是可以看作为对偶研究任务，彼此联系紧密。基于特征的图像编辑任务即为了进一步探索图像特征与语义内容的对应关联，实现可控的图像生成。

第三章　人脸编辑下的点云对齐研究

3.1　点云特征对齐及相关任务

依据图1-1中的人脸编辑的处理流程，点云特征对齐用于解决人脸对齐的问题，是人脸图像预处理的关键步骤。目前，主流研究主要关注正脸图像，且人脸大都符合标准的自然姿态。由于不同姿态角度下的人脸会显著增加下游任务的处理难度，标准且统一的人脸姿态在解决人脸编辑问题中至关重要。人脸编辑主要围绕标准姿态的人脸图像展开。不同姿态、表情、光照下的人脸会影响网络模型对图像的内部编码，难以实现高质量的图像编辑效果。为了端到端地实现人脸图像的特征编解码，人脸对齐通过标准化输入的人脸图像的姿态及表情，来解决复杂的人脸图像问题。当前，人脸对齐的解决方法大都基于特征点的对齐过程。然而，在二维图像中，受到人脸表情及姿态的影响，所提取的特征缺少鲁棒性，难以精确地将其对齐到标准姿态正脸模型中。针对此问题，我们试图将二维平面的人脸反投影到三维空间中，利用人脸的点云特征点，解决局部点云对齐并完成对人脸图像的对齐。本章围绕局部点云对齐问题展开研究，其可用于人脸编辑中的图像对齐过程。

点云对齐是解决图像编码和解码的特征对齐的过程。基于 StyleGAN 的人脸生成模型，将 Style 编码看作为一种人脸特征编码。借鉴第4章中无监督训练图像特征编码的学习方法，利用预训练的 StyleGAN 人脸生成模型，将学习人脸图像特征编码器。人脸特征编码器通过训练学习到 Style 特征编码空间的映射。目前，Edit4Edit 方法 [63] 等工作解决了人脸 Style 特征编码网络和解码生成网络的特征对齐。然而，在基于 Style 编码的解码和生成的人脸图像中，其主要围绕标准姿态下的人脸图像展开。对于真实的人脸图像而言，需要对其进行人脸定位及对齐等预处理，才能对合适人脸进行 Style 特征编解码的过程。因此，对于输入的人脸图像，需要将其与标准姿态人脸进行对齐。目前，二维图像的人脸对齐问题依靠人脸关键点检测等技术，已经大致得到了解决。取而

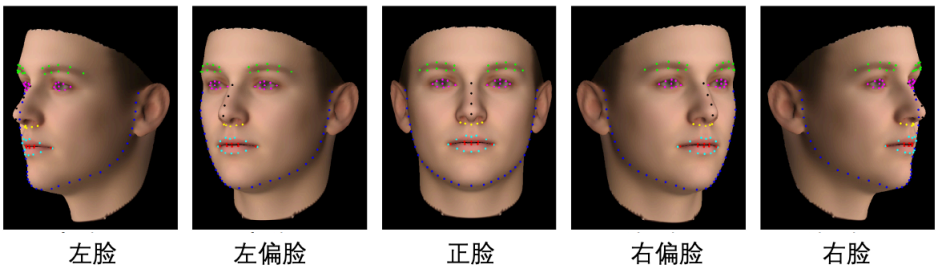


图 3-1: WFLW 数据集 [65] 中同一个人不同角度的人脸数据。

代之，三维人脸模型的对齐成为研究的热点问题。三维人脸模型的对齐解决方法能够有效解决人脸对齐问题。如图 3-1所示，WFLW 数据集 [65] 提出了基于三维人脸关键点的人脸对齐问题。在图 3-1中，正脸实现了后续的人脸特征编解码的统一。左偏脸和右偏脸可通过进行对齐转正。左脸和右脸难以直接对齐到正脸，不利于人脸编辑的研究。

点云对齐旨在找到一种刚性变换来对齐多个局部点云。传统的优化方法试图迭代计算变换矩阵，通过最小化对齐的几何误差来解决局部点云的对齐。具体的说，传统的对齐方法旨通过迭代最近点 (ICP) 等优化方法来生成旋转矩阵和平移向量。然而，由于 ICP 仅限于具有小旋转的刚性变换，大都优化方法局限于局部对齐，难以处理大角度的全局对齐。由于优化过程受局部极小值的影响，全局对齐问题一直困扰着迭代对齐方法。近年来，随着 3D 深度学习技术的发展，越来越多的基于深度学习的方法被提出用以解决优化方法的缺陷。DCP [19] 实现了端到端的局部点云对齐，但只在 ModelNet 数据集 [20] 上表现不错。DeepGMR [21] 采用混合概率分布进行点匹配，其参数由神经网络估计，并在 ICL-NUIM[22] 得到了验证。Deep Global Registration (DGR)[23] 提出了一个可区分的成对注册框架，包括特征对应预测、姿势估计和姿势细化。基于深度学习的对齐方法通常比迭代全局对齐方法运行速度快一个数量级，并且提高了全局对齐精度。深度学习对齐方法是对传统迭代优化对齐方法造成了巨大冲击。具有竞争力的迭代对齐方法是亟需新的研究成果。

本章立足于优化对齐方法，提出了混合优化的局部点云对齐方法（Hybrid Optimization method with Unconstrained Variables，HOUV）。HOUV 直接定义了一个平滑变换矩阵，通过优化其实现成对局部点云的对齐。考虑对齐过程中的局部点匹配，倒角距离（Chamfer Distance，CD）通常用于测量一对局部点云的距离。基于倒角距离，我们提出了两种变体损失。1) **局部 CD 损失**解决了成

对部分点云中的不完全匹配。2) **投影 CD 损失**实现全局匹配提高了对齐精度。同时，在优化变换矩阵的过程中，通常的优化方法会限制旋转角的平移距离。一般的优化方法常被看作为带约束变量的优化过程。在 HOUV 中，采用映射函数将约束变量替换为无约束变量，以实现平滑优化。受分支定界法的启发，将旋转平移变换 SE(3) 空间划分为几个独立的子空间。根据变换的范围，HOUV 通过引入边界来处理不同的子空间。我们提出了一种通过随机初始化多组变量的批处理策略和一种可编程优化策略来缓解局部最小值问题。对比基于深度学习的方法，HOUV 展示了如下方面的优点：1) **无监督的**，即 HOUV 可以直接应用于任何点云的配准，而无需在训练数据集上学习。2) **可编程性**，即优化策略是可编程的以适应点云样本或数据集。HOUV 方法可以控制超参数来处理不同程度的重叠的局部点云。3）**高对齐精度**。在伪造和真实的局部点云数据集上，HOUV 展示了最先进的对齐结果。对比基于深度学习的方法，HOUV 方法体现了传统迭代优化方法仍有巨大的研究潜力。本章的主要研究内容如下所示：

- 人脸对齐解决人脸编辑问题中复杂人脸图像的特征编码问题，提高了人脸特征稳定性。
- 针对人脸对齐，提出了混合优化的局部点云对齐方法，提高了点云的对齐精度和速度。
- 实验验证了提出的 HOUV 方法在点云对齐的有效性，利用点云完成了对人脸图像的对齐。

3.2 局部点云对齐问题及解决方法

首先，形式化地定义局部点云配准问题。然后，引入无约束变量的混合优化对齐方法（Hybrid Optimization with Unconstrained Variables method ，HOUV）分别构造旋转矩阵和平移向量。我们提出的混合损失包含倒角距离（Chamfer Distance，CD）的两种变体，分别处理局部和全局点匹配。最后，为了避免随机初始化导致的优化不稳定，本文采用批处理策略来提高 HOUV 的稳定性。同时，考虑到对齐变换的子空间，HOUV 通过引入边界来缩小搜索空间 HOUV 可利用局部点云的先验知识来调整配准的边界。

3.2.1 局部点云特征对齐问题

在 3D 坐标系中，我们有一个待观察的刚性物体和一个可移动的观察点。点云 $P = \{\mathbf{p}_i \in \mathbb{R}^3 \mid i = 1, \ldots, N\}$ 和 $Q = \{\mathbf{q}_j \in \mathbb{R}^3 \mid j = 1, \ldots, M\}$ 分别是从视点 $\mathbf{v}_p$ 和 $\mathbf{v}_q$ 观察得到。其中 N 和 M 表示每个点云中的点数。假设观察点 $\mathbf{v}_q$ 由 $\mathbf{v}_p$ 通过未知的刚性旋转矩阵 $\mathbf{R}_{view}$ 和平移向量 $\mathbf{t}_{view}$ 得到，形式化定义为：

$$\mathbf{v}_q = \mathbf{R}_{view} \cdot \mathbf{v}_p + \mathbf{t}_{view}. \tag{3-1}$$

如果观察点 $\{\mathbf{R}_{view}, \mathbf{t}_{view}\}$ 已知，可以对齐这两部分的局部点云以获得融合点云，用以更好地表示物体。然而，当观察点 $\mathbf{v}_p$ 和 $\mathbf{v}_q$ 通常情况下是未知的，我们只能直接预测点云对齐结果 $\{\mathbf{R}_{pq}, \mathbf{t}_{pq}\}$。

点云对齐旨在估计刚性变换 $\{\mathbf{R}_{pq}, \mathbf{t}_{pq}\}$ 对齐源局部点云 P 和目标点云 Q，其中 $\mathbf{R}_{pq} \in SO(3), \mathbf{t}_{pq} \in \mathbb{R}^3$。当成观察点 $\mathbf{v}_p$ 和 $\mathbf{v}_q$ 具有重叠区域时，局部点云 P 和 Q 可进行局部匹配点的对齐。局部点云的变换矩阵 $\{\mathbf{R}_{pq}, \mathbf{t}_{pq}\}$ 可以通过匹配点进行估计，即

$$(\mathbf{R}_{pq}, \mathbf{t}_{pq}) = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{(\mathbf{p}_k, \mathbf{q}_k) \in CP} (\|\mathbf{R} \cdot \mathbf{p}_k + \mathbf{t} - \mathbf{q}_k\|^2), \tag{3-2}$$

其中特征点集 CP 由局部点云 P 和 Q 的匹配点组成。匹配点表示成对粒子点云中的重叠区域。在两个局部点云对中，找到具有正确匹配点的子集不是一个简单的问题。

3.2.2 无约束变量的变换矩阵

变换矩阵由一个旋转矩阵和平移向量组成。根据 Rodrigues 的公式 [66]，任何三维旋转矩阵都可以由其轴 $\mathbf{v}$ 和角度 θ 定义。旋转矩阵 $\mathbf{R}$ 可以用旋转轴和角度来表示，即

$$\mathbf{R} = \cos(\theta)\mathbf{I} + (1 - \cos(\theta))\mathbf{v}\mathbf{v}^T + \sin(\theta)\mathbf{v}^\wedge, \tag{3-3}$$

其中 $\|\mathbf{v}\| = 1, \theta \in [0, \pi]$ 和

$$\mathbf{v}^\wedge = \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix}. \tag{3-4}$$

向量 $\mathbf{v} = (v_x, v_y, v_z)$ 可以是任何单位向量。点云对齐旨在从点云 P 估计到点云 Q 的变换矩阵 $\{\mathbf{R}_{pq}, \mathbf{t}_{pq}\}$。定义变量 $\{\mathbf{v}_{pq}, \theta_{pq}\}$ 以通过式 (3-3) 生成旋转矩阵 $\mathbf{R}_{pq}$。由于变量 $\mathbf{v}$ 的模在优化过程中被限制为 1，采用规范化的变量 $\mathbf{v}_{pq}$，用以解决该约束问题。

$$\mathbf{v} = \frac{\mathbf{v}_{pq}}{\|\mathbf{v}_{pq}\|} \tag{3-5}$$

由于 cos(·) 是一个对称的周期函数，可以直接在 $\mathbb{R}$ 上优化变量 θ_{pq}。因此，采用无约束变量 $\{\mathbf{v}_{pq}, \theta_{pq}\}$ 代替了约束变量 $\mathbf{R}_{pq}$。

然后，解决局部点云对齐的平移问题，主要关注平移方向和平移距离。平移方向通常被视为一个单位向量。平移向量 $\mathbf{t}$ 可以用平移方向 $\mathbf{u}$ 和距离 d 表示为

$$\mathbf{t} = d\mathbf{u}. \tag{3-6}$$

为了缩小平移的优化空间，我们限制平移距离：$d <= d_{max}$（常数 d_{max} 表示平移距离的上限）。此约束在优化问题中称为“框约束”。通常，解决框约束问题有三种不同的方法 [67]，即

1) 执行标准梯度下降，然后裁剪所有变量以满足约束。
2) 将约束函数合并到要最小化的目标函数，并求解广义拉格朗日函数。
3) 引入无约束变量代替原来的约束变量来优化目标函数。

当使用第一种方法，待优化变量往往集中在边界上，使得问题得不到充分解决。第二种方法根据拉格朗日对偶性要求目标函数是凸的。因此，我们采用第三种方法，使用映射函数将受约束的 d 替换为不受约束的 d_{pq}。映射函数根据平移距离将变量 d_{pq} 扩展到 $\mathbb{R}$ 空间。利用无约束变量 $\{d_{pq}, \mathbf{u}_{pq}\}$ 和映射函数 sigmoid(·) 重构式 (3-6) ，表示为

$$\mathbf{t}_{pq} = d_{max}\text{sigmoid}(d_{pq})\frac{\mathbf{u}_{pq}}{\|\mathbf{u}_{pq}\|}. \tag{3-7}$$

与此类似，也可以使用 sin(·) 作为映射函数，表示为

$$\mathbf{t}_{pq} = 1/2(d_{max} + d_{max}\sin(d_{pq}))\frac{\mathbf{u}_{pq}}{\|\mathbf{u}_{pq}\|}. \tag{3-8}$$

不同的映射函数在生成平移距离具有不同的偏好。映射函数可根据数据集中平移距离的分布来灵活进行选择。

3.2.3 混合局部和全局对齐损失

为了处理 (3-2) 中的对齐问题，我们引入倒角距离 CD 作为两个局部点云的评估指标。原始 CD 将每个点都考虑在内，以在另一个点云中找到最近的点。对于成对的局部点云，CD 仅在与重叠区域相关的对应点上合理工作。在本文中，采用 CD 损失的两种变体用以进行局部和全局的点云对齐。

局部 CD 损失。由于一对局部点云只与对应的匹配点有关，因此在整个局部点云上计算 CD 损失是不合理的。局部 CD 损失集中在局部匹配点上，解决不同视角下局部点云的偏差。局部 CD 损失定义为

$$\begin{aligned}\mathcal{L}_{CD_{local}}(P,Q,\alpha) = &\frac{1}{|P_\alpha|}\sum_{\mathbf{p}\in P_\alpha}\min_{\mathbf{q}\in Q}\|\mathbf{p}-\mathbf{q}\|^2 \\ &+\frac{1}{|Q_\alpha|}\sum_{\mathbf{q}\in Q_\alpha}\min_{\mathbf{p}\in P}\|\mathbf{q}-\mathbf{p}\|^2,\end{aligned} \tag{3-9}$$

其中 P_α 和 Q_α 分别是 P 和 Q 的子集。超参数 α 与局部点云重叠的比例有关。点集 P_α 是按如下过程从局部点云 P 中选择的：(1). 在每一轮中，局部点云 P 中的点 p 依据 $\min_{\mathbf{q}\in Q}\|\mathbf{p}-\mathbf{q}\|^2$ 选取，并从点集 P 中删除该点。(2). 重复过程（1）并停止，直到选择了 $\alpha|P|$ 个点。子集 Q_α 以相同的方式生成。局部 CD 损失通过调整 α 来控制从原始点云到目标点云的匹配点数量。通过超参数 α（只需要调整一个超参数），局部 CD 损失可以用于不同重叠区域的各种对齐任务。

投影 CD 损失。局部 CD 损失只关注局部对齐而忽略了全局点匹配。当重叠中的对应点多于具有固定超参数 α 的选定点时，局部 CD 损失难以应对这种情况。将三维空间中的点投影到二维平面上，保留轮廓特征进行全局配准。基于这一观察，我们将点云投影到 x-y、y-z、x-z 平面上，并将投影的 CD 损失计算为

$$\begin{aligned}\mathcal{L}_{CD_{uv}}(P,Q) = &\frac{1}{|P|}\sum_{\mathbf{p}\in P}\min_{\mathbf{q}\in Q}\|\mathbf{p}_{uv}-\mathbf{q}_{uv}\|^2 \\ &+\frac{1}{|Q|}\sum_{\mathbf{q}\in Q}\min_{\mathbf{p}\in P}\|\mathbf{q}_{uv}-\mathbf{p}_{uv}\|^2,\end{aligned} \tag{3-10}$$

其中 uv 表示投影平面，$\mathbf{p}_{uv}$ 和 $\mathbf{q}_{uv}$ 是投影的方向向量。

算法 3.1 原始的无约束变量对齐方法 ($HOUV_{ori}$)

Input：成对的局部点云 P 和 Q。
Parameter：组大小 g，学习率 l，迭代轮数 $iter$，超参数 α 和 β。
Output：具有无约束变量的旋转 $\mathbf{R}_{pq}$ 和变换 $\mathbf{t}_{pq}$ 矩阵。

```
依据组大小 g，随机初始化变量组 {v_pq, θ, u_pq, d_pq}。
for 对于每一组初始化变量 do
    while 未达到迭代轮数 iter do
        依据式 (3-3) 和 (3-8)，计算旋转 R_pq 和平移 t_pq。
        依据式 (3-11)，计算混合优化损失。
        执行反向传播以计算变量相对于混合损失的梯度。
        按照梯度下降法更新变量。
    end while
end for
选择局部 CD 损失最小的一组，作为最终预测结果。
return {v_pq, θ, u_pq, d_pq, R_pq, t_pq}.
```

混合优化损失。混合优化损失由局部 CD 和投影 CD 损失组成，定义为

$$\begin{aligned}\mathcal{L}_{hybrid}(P,Q) = &\mathcal{L}_{CD_{local}}(P,Q,\alpha)\\ &+\beta(\mathcal{L}_{CD_{xy}}(P,Q)+\mathcal{L}_{CD_{yz}}(P,Q)\\ &+\mathcal{L}_{CD_{xz}}(P,Q)),\end{aligned} \tag{3-11}$$

其中 β 用于平衡局部和全局对齐损失。基于混合优化损失，我们通过梯度下降迭代更新不受限制的变量。

3.2.4 无约束变量的对齐优化方法

由通过梯度下降预测刚性变换是一个非凸优化问题。我们初始化多个实例并选取最优结果，来解决 HOUV 优化过程中的局部极值问题。对于每个成对的部分点云，将旋转轴 $\mathbf{v}_{pq}$ 和角度 θ 初始化为 32 组。利用 GPU 上的并行计算，这些组变量可同时得到优化。最后，选择局部 CD 损失最小的预测变换作为最终的局部点云对齐结果。假设成对部分点云的最佳配准满足这些组中的最小局部 CD 损失。算法 4.1 详细展示了提出的原始 HOUV 方法。

HOUV 不仅可以解决整个 SE(3) 空间的配准问题，还可以通过引入边界来调整搜索子空间。边界来自对局部点云对齐问题的分析，并获取变换的先验范围。当整个 $SE(3)$ 空间被划分为若干个子空间时，HOUV 可以对每个子空间进行细粒度的优化。该策略提高了推理对齐的速度和准确性，部分缓解了点云物

算法 3.2 基于先验变换子空间的无约束变量对齐方法 ($HOUV_{pri}$)

Input：成对的局部点云 P 和 Q。
Parameter：组大小 g，学习率 l，迭代轮数 $iter$，超参数 α 和 β，变换子空间边界 $boundry$。
Output：具有无约束变量的旋转 $\mathbf{R}_{pq}$ 和变换 $\mathbf{t}_{pq}$ 矩阵。

依据组大小 g，随机初始化变量组 $\{\mathbf{v}_{pq}, \theta, \mathbf{u}_{pq}, d_{pq}\}$。
for 对于每一组初始化变量 **do**
  while 未达到迭代轮数 $iter$ **do**
    根据式 (3-12)，利用边界条件 boundry 设置最大旋转角度和平移距离。
    依据式 (3-3) 和 (3-8)，计算旋转 $\mathbf{R}_{pq}$ 和平移 $\mathbf{t}_{pq}$。
    依据式 (3-11)，计算混合优化损失。
    执行反向传播以计算变量相对于混合损失的梯度。
    按照梯度下降法更新变量。
  end while
end for
选择局部 CD 损失最小的一组，作为最终预测结果。
return $\{\mathbf{v}_{pq}, \theta, \mathbf{u}_{pq}, d_{pq}, \mathbf{R}_{pq}, \mathbf{t}_{pq}\}$.

体的对称问题。对于对称物体，在旋转角度上加上 π 或 $\pi/2$ 度时，局部点云仍然对齐。通过指定搜索子空间，HOUV 可以解决上述问题。类似于 (3-7)，我们将旋转角度 θ_{pq} 限制为 $[l, r]$，可表示为

$$\theta_{pri} = l + (r - l)\sin(\theta_{pq}), \tag{3-12}$$

其中 l 和 r 分别是旋转角度的下限和上限。我们应用该优化策略的 HOUV 可以灵活地适应不同的变换分布。算法 3.2 中总结了利用点云数据先验的 HOUV 方法。

3.3 实验设计与分析

本章实验主要解决人脸图像的标准姿态对齐问题，用于减轻后续人脸编辑的难度。通常，二维的人脸图像常可以通过局部关键点进行对齐。然而，平面上的局部关键点易受到人脸表情和姿态的影响以造成定位的误差。因此，本章利用最新的三维关键点预测技术预测人脸图像的空间关键点，并提出了高效的点云对齐方法，实现了人脸图像的标准姿态对齐处理。我们将在多类点云模型上验证提出方法的有效性，并将其应用到人脸图像的对齐处理。

3.3.1 数据集与评价指标

数据集：我们在三个合成点云数据集上评估了所提出的局部点云对齐方法。同时，该方法也被应用于真实环境获取的局部点云对齐。数据集的具体内容如下：

1. **Multi-View Partial Point Cloud (MVP)** 数据集 [68] 是一个高质量的多视图局部点云数据集，其中包含超过 100,000 个高质量扫描模型。对于每个 CAD 模型，部分点云是从 MVP 中的 26 个均匀分布的相机位姿中获取的。
2. **ModelNet** 数据集 [20] 包括来自 40 个类别的 12311 个网格化 CAD 模型。对于数据集中的每个对象，该数据集随机抽取 1024 个点作为源局部点云 Q，然后对 Q 进行随机变换，并通过打乱这些点来获得未对齐的局部点云 P。同时，ModelNet 向局部点云 Q 添加来自 $\mathcal{N}(0, 0.01)$ 的高斯噪声以模拟真实世界的噪声。
3. **ICL-NUIM** 是由伦敦帝国理工学院（The Imperial College London）和爱尔兰国立大学梅努斯（The National University of Ireland Maynooth ）创建的数据集 [22]，用于评估视觉里程计、3D 重建和 SLAM 算法。他们使用从增强型 ICL-NUIM 数据集中的 RGB-D 扫描派生的局部点云，然后用真实的传感器噪声和输入对之间的任意转换来增强该局部点云。
4. **3DMatch**[69] 由来自各种真实世界场景的 3D 点云对组成，具有从 RGB-D 重建管道估算的地面实况转换。我们在测试数据集上比较了我们的无监督方法和其他方法，其中生成的部分点云对至少有 30% 重叠 [70]。

评价指标：为了分析在上述局部点云数据的对齐精度，我们选择各向同性旋转误差 Rot_{err}、各向同性平移误差 $Trans_{err}$ 和均方误差 (Mean Square Error，MSE) 作为度量指标。将两个局部点云的预测对齐结果表示为 $\mathbf{T}_{pq} = \{\mathbf{R}_{pq}, \mathbf{t}_{pq}\}$，并将其与真实对齐结果 $\mathbf{T}_{gr} = \{\mathbf{R}_{gr}, \mathbf{t}_{gr}\}$ 进行对比。其中，Rot_{err} 计算 $\mathbf{R}_{pq}$ 和 $\mathbf{R}_{gr}$ 旋转之间的角距离为

$$Rot_{err} = \frac{180}{\pi} \arccos\left(\frac{\left(\text{Trace}\left(\mathbf{R}_{pq}{\mathbf{R}_{gr}}^{T}\right) - 1\right)}{2}\right). \tag{3-13}$$

为了分析方便，我们采用角度表示预测旋转角与真实旋转角的误差。$Trans_{err}$

和 MSE 使用向量间的 L_2 平移距离和局部点云的对应点进行评估。MSE 定义为

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{T}_{pq}(\mathbf{p}_i) - \mathbf{T}_{gr}(\mathbf{p}_i)\right\|^2, \tag{3-14}$$

其中 $\mathbf{T}_{pq}(\mathbf{p}_i)$ 表示为 $\mathbf{R}_{pq}\mathbf{p}_i + \mathbf{t}_{pq}$ 和 $\mathbf{T}_{gr}(\mathbf{p}_i)$ 表示为 $\mathbf{R}_{gr}\mathbf{p}_i + \mathbf{t}_{gr}$。

3.3.2 对比方法及相关设置

本文测试了 ICP[71]、Fast Global Registration (FGR)[72] 作为基本的迭代优化方法。ICP 和 FGR 由 Open3D 程序库实现[①]。Euler 方法是一种源自 $HOUV_{pri}$ 的约束优化方法，在消融实验作为对比方法。此外，我们还对比了三种基于深度学习的点云对齐方法，分别为 DCP[19]、IDAM[73] 和 DeepGMR[21]。这三种方法作为 MVP 对齐比赛的基准对比方法。RGM[74] 是 ModelNet 上当前最先进的方法。HOUV 默认表示为 $HOUV_{ori}$。我们为 HOUV 设置超参数 $\alpha = 0.5$ 和 $\beta = 0.1$。对于 ModelNet 和 ICL-NUM 数据集，其他相关方法的结果参考了 [21]，主要包括 HGMR[75]、PointNetLK[76]、DGR[23] 和采用 RANSAC 优化的 ICP 方法 [77]。

对于迭代优化方法，将迭代次数统一设置为 500。HOUV 为一对部分点云随机初始化 32 组变量。Adam[78] 被选为默认优化器，学习率设置为 0.001。对于 MVP 数据集，还使用 $HOUV_{pri}$ 来提高在比赛中的表现。$HOUV_{pri}$ 将 180 度分为四个区间，每个区间的旋转角度为 45 度。对于 ICP 和 FGR 方法，还为每种方法随机初始化 32 组进行测试。最后选择具有最高适配度的变换矩阵作为预测结果。对于基于深度学习的对齐方法，DCP、IDAM 和 DeepGMR 等方法均通过数据增强进行了预训练。

3.3.3 局部点云的对齐实验

局部点云的对齐结果：HOUV 方法在合成数据集上表现出优越的性能。如表 3-1 所示，在 MVP 数据集上 HOUV 在各项性能指标上优于其他方法。虽然基于深度学习的方法优于 ICP 和 FGR 等传统优化方法，但它们的性能弱于本文提出的迭代方法。图 3-2 使用 HOUV 方法后，可视化展示了成对局部点云的对齐结果。我们调查了传统方法例如 ICP 方法旋转误差大的原因。对于 MVP

[①] http://www.open3d.org/

表 3-1: HOUV 方法和其他对齐方法在 MVP 数据集上的结果。

Method	ICP	FGR	GO-ICP	DCP	IDAM	DeepGMR	RGM	Euler	DGR	HOUV
Rot_{err}	35.87	30.47	56.76	26.27	22.87	49.95	17.4	5.63	20.89	**3.05**
$Trans_{err}$	0.14	0.15	0.24	0.23	0.23	0.38	0.10	0.03	0.21	**0.02**
MSE	0.68	0.60	0.79	0.64	0.62	0.70	0.42	0.13	0.57	**0.07**

表 3-2: 当召回率设置为 0.2，ModelNet 和 ICL-NUIM 数据集上的 RMSE 指标。

	ModelNet clean		ModelNet noisy		ModelNet unseen		ICL-NUIM	
	RMSE	Re@0.2	RMSE	Re@0.2	RMSE	Re@0.2	RMSE	Re@0.2
ICP	0.53	0.41	0.53	0.41	0.59	0.32	1.16	0.27
FGR	0.19	0.79	0.19	0.79	0.23	0.75	0.15	0.87
HGMR	0.52	0.44	0.52	0.45	0.54	0.43	0.72	0.50
ICP + RANSAC	0.08	0.91	0.42	0.49	0.30	0.67	0.17	0.84
PointNetLK	0.51	0.44	0.56	0.38	0.68	0.13	1.29	0.08
DCP	0.02	0.99	0.08	0.94	0.34	0.54	0.64	0.16
DeepGMR	**<0.01**	0.99	**<0.01**	0.99	0.01	0.99	0.07	0.99
RGM	**<0.01**	1.0	**<0.01**	0.99	**<0.01**	0.99	0.04	0.99
DGR	**<0.01**	0.99	**<0.01**	0.99	0.02	0.99	0.05	0.99
HOUV	**<0.01**	0.99	**<0.01**	0.99	**<0.01**	0.99	**0.01**	0.99

数据集，大多数局部点云具有对称结构，如图 3-2 所示。当在成对局部点云上贪婪地应用倒角距离时，配准结果将尽可能重叠，但忽略相似结构的存在。当从 32 组中选取一组时，我们选择最高的适应度作为最终的预测结果。当成对部分点云中重叠的比例小于超参数 α 时，该策略会失效。这可能会导致部分对齐方法产生额外的 180 或 90 的旋转误差。

对于 ModelNet 和 ICL-NUIM 数据集，表 3-2展示了 HOUV 等方法的结果，其结果参考 [21]。ModelNet 和 ICL-NUIM 可以通过许多对齐方法有效地解决。因此，与 RGM 等其他方法相比，HOUV 没有表现出显着的改进（已将近 100% 的对齐精度）。尽管实验结果有限，但 HOUV 方法在多个合成数据集上验证了良好的泛化性能。

HOUV 还在逼真的 3DMatch 数据集中的各种场景中得到了验证。表 3-3 展示了 3DMatch 的对齐精度。该实验是在具有 5 cm 体素的下采样点云上实现的。我们只在 3DMatch 上评估 HOUV 方法，其他结果来自 [23]。如表 3-3所

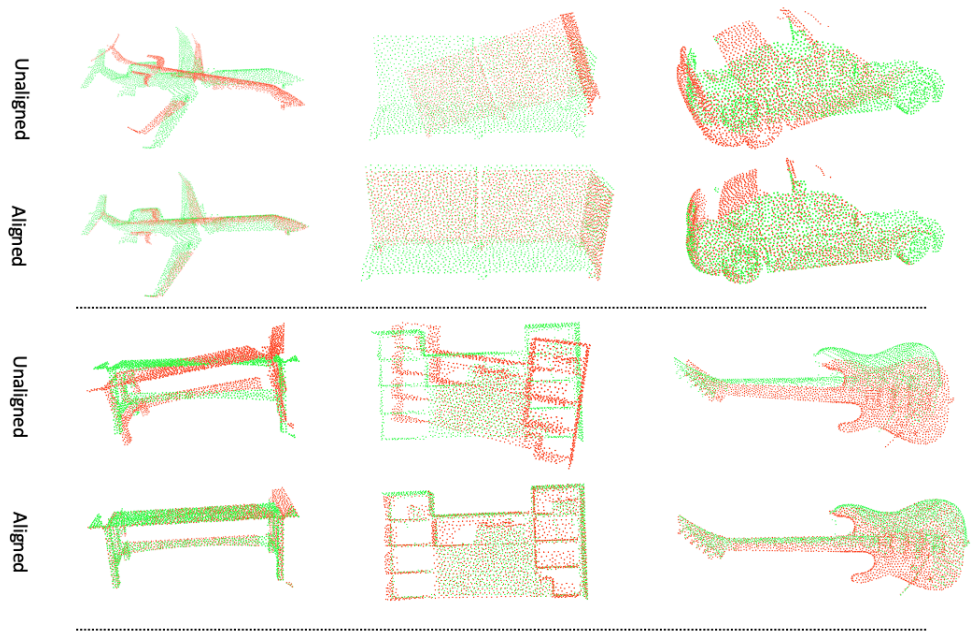


图 3-2: 在 MVP 数据集上使用 HOUV 方法预测的对齐结果。每种颜色代表一个局部点云数据。红色样本代表局部点云 P。绿色样本代表局部点云 Q。在图中的每一行中，我们展示了执行预测对齐结果后的局部点云 P。

表 3-3: HOUV 方法和其他对齐方法在真实的 3DMatch 数据集上的评价指标。

Method	ICP	FGR	GO-ICP	DCP	DGR	HOUV
Rot_{err}	8.25	4.08	5.38	8.42	2.43	3.42
$Trans_{err}$(cm)	18.1	10.6	14.7	21.4	7.34	7.77
Recall	6.04%	42.7%	22.9%	3.22%	91.3%	60%

示，HOUV 优于经典的对齐方法。同时，我们也调查了 HOUV 性能略逊于 DGR 方法的原因。首先，与 DGR 相比，HOUV 是一种无监督方法。DGR 在训练数据集中学习相似场景的先验对象特征。其次，3DMatch 数据集中有广泛分布的转换和重叠区域。HOUV 方法不能用单一固定的超参数 α 拟合所有成对的点云。HOUV 可以调整超参数 α 来对齐那些失败的局部点云对。图 3-3 可视化展示了 3DMatch 在不同场景下的对齐结果。

人脸图像的局部点云对齐结果：为了实现二维平面内人脸图像的对齐，即将任意人脸姿态表情与标准人脸模版对齐，我们首先利用人脸图像预测出人脸

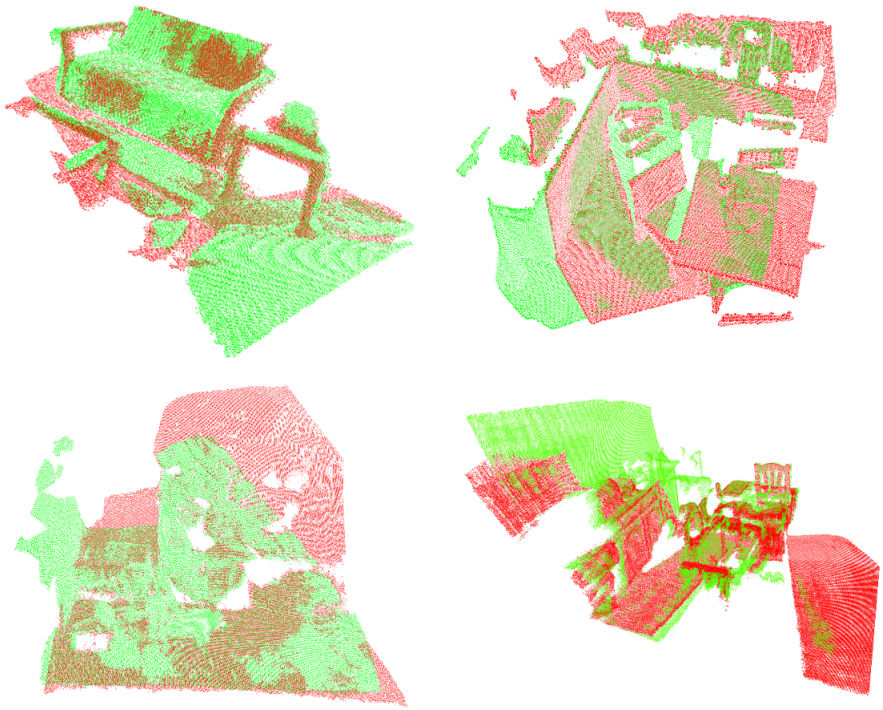


图 3-3: HOUV 在 3DMatch 中的测试场景上关于成对局部点云的可视化对齐结果。其中，红色样本代表点云 P。绿色样本代表点云 Q。

轮廓及边缘的关键点局部点云数据。我们采用 Bulat 等人 [79] 的工作，运用人脸检测模型定位人脸图像并预测 68 个关于人脸的三维关键点坐标，如图 3-4所示。这些关键分布于人脸轮廓及五官，高效地表达了人脸轮廓信息。其次，基于图像中预测的人脸关键局部点云，我们应用提出的方法 HOUV 对不同人脸表情和姿态的人脸进行了对齐。如图 3-5所示，局部点云对齐的 HOUV 方法可被应用于二维平面内的人脸图像对齐。

3.3.4 消融实验及超参数分析

在表 3-4 中，消融实验研究了 HOUV 的不同部分对点云对齐的影响。对比表 3-4 中第一行和第四行的结果，投影 CD 损失提高了成对局部点云的旋转精度。对比第二行和第四行的结果，HOUV 通过将变量 d 替换为无约束变量 d_{pq} 来减少平移误差。使用非线性映射函数的无约束变量 d_{pq} 减少了沿平移方向的平移误差。对比第三行和第四行的结果，通过为旋转角度划分更细粒度的间隔，提高了 HOUV 的性能。由于多 GPU 并行化，当采用这些策略时，推理时

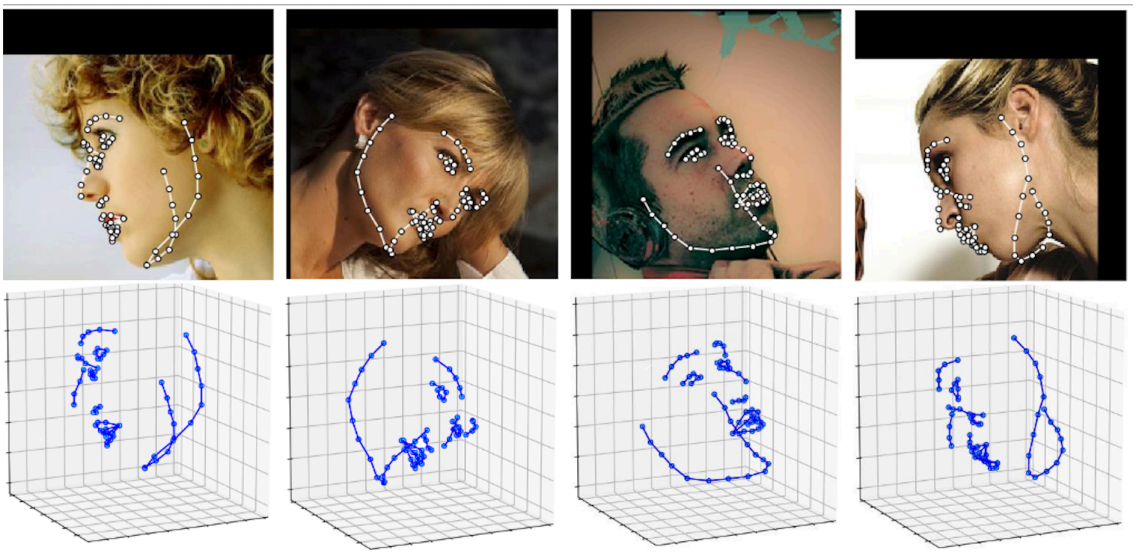


图 3-4: 二维图像上预测的三维关键点云可视化结果。

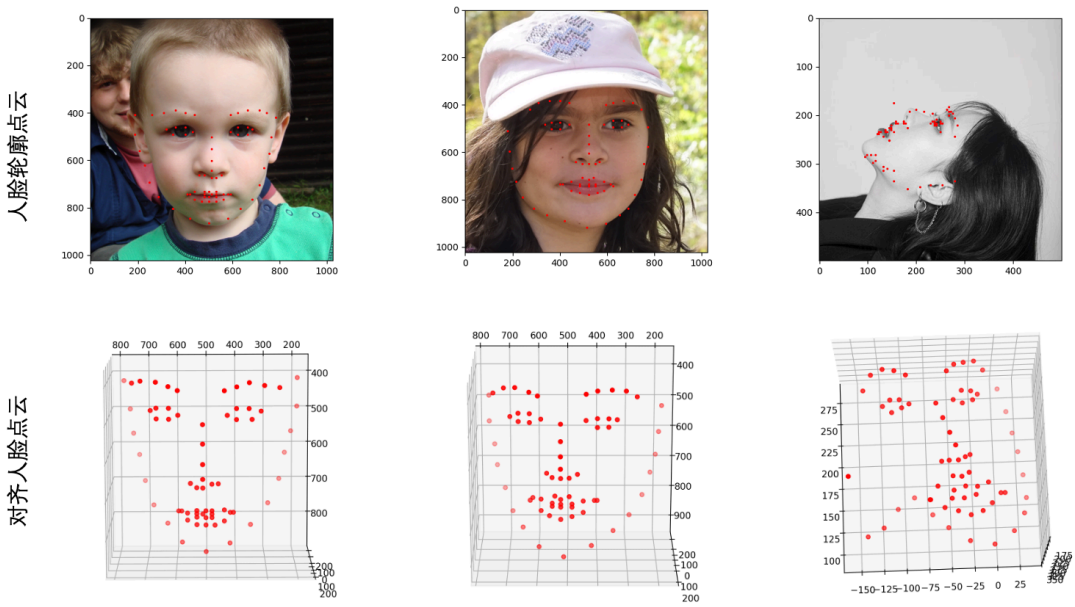


图 3-5: 利用人脸轮廓关键点的局部点云对齐可视化结果。

间不会随着划分的子空间线性增加。

图 3-6 可视化了没有投影 CD 损失的大旋转误差的结果。在此图中，我们分析了具有局部 CD 损失对 HOUV 方法在对齐精度上的作用。我们执行 HOUV 方法预测变换矩阵以对齐局部点云 P 与 Q 。在图 3-6 的第二列中，HOUV 在局部区域拟合得很好，但在仅使用局部 CD 损失时在全局的对齐上失败。原因

表 3-4: HOUV 方法在 MVP 数据集上的消融实验结果。

Local CD	Projected CD	Unconstrained T	Strategies	Rot Error	Trans Error	MSE
✓		✓	✓	3.4654	0.0229	0.0799
✓	✓		✓	3.2349	0.0291	0.0795
✓	✓	✓		3.8390	0.0238	0.0734
✓	✓	✓	✓	2.9987	0.0213	0.0719

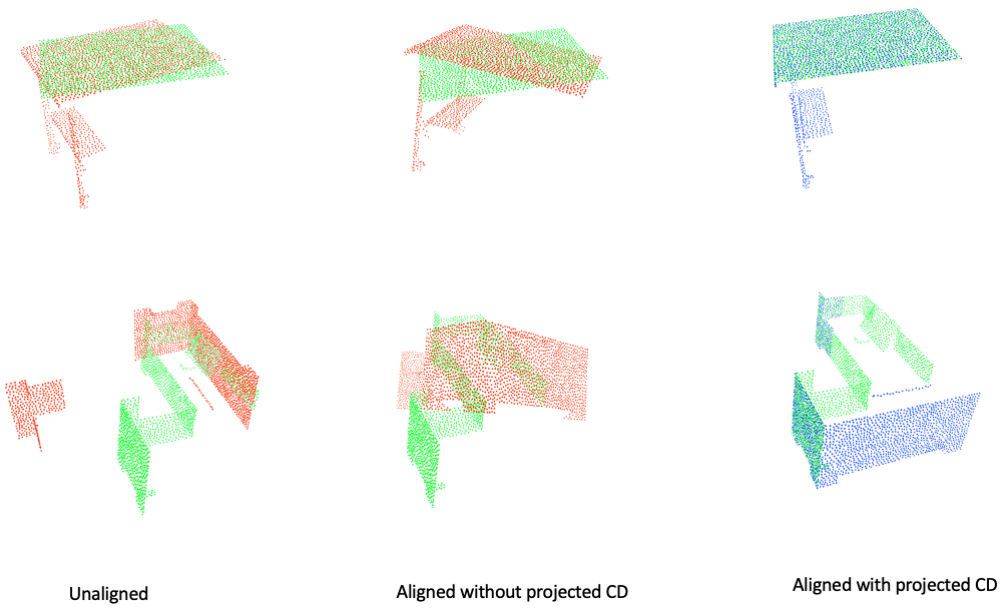


图 3-6: 验证投影 CD 损失对于 HOUV 方法对齐局部点云作用的可视化结果。红色样本代表没有利用投影 CD 损失的 HOUV 方法对齐的局部点云 P。绿色样本代表目标局部点云 Q。蓝色样本代表利用投影 CD 损失的 HOUV 方法对齐的局部点云 P。

主要是在于局部 CD 损失设置超参数 α 为 0.5 过小。如图 3-6 的第三列所示，投影 CD 损失可以解决重叠比例大于超参数 α 的样本。投影 CD 损失可以弥补局部 CD 损失不足，提升对齐局部点云的精度。

3.4 本章小结

本章主要面向人脸编辑问题下的图像对齐研究，并围绕成对局部点云对齐问题，提出了无约束变量的混合优化对齐方法。我们提出的 HOUV 迭代优化方法，能有效解决局部点云的对齐。在构建变换矩阵时，HOUV 将无约束变量代替原本的约束变量，简化了优化过程。为了实现迭代优化，HOUV 提出局部

CD 损失和投影 CD 损失，用于无约束变量的优化。其中，局部 CD 损失解决局部特征对齐，投影 CD 损失解决全局特征对齐。局部点云特征的对齐研究可用于人脸图像的对齐，是人脸编辑中预处理的关键步骤。后续的人脸编辑过程中，将主要面向标准姿态的人脸图像进行研究。人脸图像对齐的预处理过程简化了后续人脸特征编码、解码和编辑的难度。

第四章　人脸编辑下的特征编码研究

4.1　图像特征编码及学习问题

从图像中提取有语义信息的过程，称为图像语义编码，也称为图像特征编码。不同于手工提取图像特征的研究内容，如角点、SIFT 关键点 [80] 等特征，本节主要面向神经网络模型下图像特征编码的学习问题。目前，深度网络模型大多采用端到端的方式来解决具体的图像任务，提取图像特征的过程常被隐含在各种任务中。并且，由于神经网络模型对学习到的特征编码缺乏可解释性，使得衡量不同网络模型学习的特征编码的好坏成为棘手问题。目前，通过评估在下游具体任务的性能，可以从侧面分析图像特征编码的优劣。图像特征与各类图像任务息息相关。例如，图像分类模型常被看作为特征提取网络和分类网络的线性连接。图像特征提取网络可以实现图像到抽象语义的特征编码，并利用分类网络对该编码进行区分。ResNet 网络，EffieientNet 网络和 ViT 网络探索了不同结构的提取网络对图像特征编码的影响。

人脸编辑离不开对网络模型内部图像特征的研究。首先，网络模型中学习的图像特征与图像语义内容存在密切联系。例如，特征编码的可解释性研究 [81] 分析了不同层级特征在边缘、纹理和目标等图像语义上的关联。其次，我们可以利用解耦的特征编码实现图像语义内容的修改。例如，GANSpace[40] 采用 PCA 方法分析了 StyleGAN 中的 Style 编码空间，将主成分特征向量用于对特征编码的扰动，并查明对应生成图像的编辑内容。因此，学习通用的图像特征编码网络有利于人脸图像编辑的处理。

基于监督学习的特征编码是主流手段。在该任务中，学习图像语义编码主要分为两个过程。首先，通过学习后的图像编码网络，提取图像中的语义内容并生成特征编码。其次，将提取的特征编码应用在众多的下游图像子任务中，验证学习的图像编码网络的性能。其中，图像编码网络成为本文研究图像编码的主要内容。不同的图像编码网络会影响特征编码的优劣，差异主要体现在网

络层的类型，组织结构和学习过程。然而，上述研究主要集中在解决手工标注的图像数据集上的分类问题。不同结构的网络模型都采用相似的有监督学习方式进行训练，利用相同的标注数据集。由于手工标注图像分类的数据集代价是巨大的，限制了模型学习的规模。例如，ImageNet[4] 数据集收集并标注了 1500 万张图片，其中主要包含有 1000 个类别。手工标注图像分类数据集的困难限制了图像分类任务的发展，也限制着图像编码网络的性能。为了解决数据标注的困境，本章从图像编码模型的学习过程出发，尝试利用无标注图像数据学习图像特征编码网络。面向无监督学习的特征编码问题，从无标注数据中提取图像特征主要面临两大挑战。第一，如何从无标注数据中挖掘监督信息用于训练图像编码网络。第二，如何评估网络提取的图像编码的优劣。为了解决上述难点，本节将图像编码的研究与图像聚类任务进行关联。首先，图像聚类任务主要解决无标注的图像集的划分，对相似图像进行聚类，符合问题一的描述。其次，图像聚类任务适合对学习到的图像编码进行比较和分析，有效解决了问题二。因此，图像聚类任务需要围绕上述的研究难点展开。其一，图像聚类研究离不开对图像编码网络的无监督学习。其二，图像聚类任务上的性能评估可以图像编码网络的性能。本文通过聚焦于解决图像聚类问题，研究无监督下图像编码网络的学习。

为了研究无标注数据的图像特征编码学习问题，本文结合了图像聚类任务并提出了深度降维嵌入聚类方法（Deep Embedded Dimensionality Reduction Clustering，DERC）[82]。DERC 将图像聚类的过程分为三个部分。第一部分，特征编码网络部分。该部分搭建合适的编码网络实现图像到语义空间的映射。当训练编码网络完成后，该特征编码被用于直接或间接聚类。第二部分，降维聚类部分。对图像特征编码进行的降维和聚类，实现初步的图像聚类。第三部分，聚类和特征编码优化。利用初步的聚类结果，优化特征编码网络学习更离散的特征编码，提高不同图像类间的区分度。同时，DERC 的核心研究是无监督下的图像编码的学习问题。本章还从学习方法角度出发，对 DERC 中的编码网络的学习方式进行了分析，并在实验环节验证了不同的学习方法对于图像编码的影响。综上所述，本章的主要内容有：

- 基于图像聚类任务，研究从无标注数据中学习图像语义特征。
- 提出了 DERC 聚类模型，有效解决相似图片的聚类问题。
- 分析了 DERC 的训练方法中存在的多种学习方法，并用于图像语义特征的学习。

- 实验验证了 DERC 方法在图像聚类中的有效性，并可视化分析了不同学习方法下学习到的图像特征。

4.2 图像特征编码及 DERC 方法

首先，为了从无标注图像中，解决特征编码的问题。将特征特征编码的学习问题，应用到具体的图像聚类任务中。特征特征编码被形式化定义在图像聚类问题中。其次，介绍了 DERC 的方法流程，并详细介绍了 DERC 中每个模块的研究内容。最后，给出了 DERC 方法中解决图像特征编码的方法。

4.2.1 图像特征编码问题

图像聚类问题中，假设有数据集 $\mathcal{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}]$ 需要将其划分为 K 类的图像簇，其中 N 为数据集中图像的数量。本章中将图像按照行进行合并表示成向量形式，即图像 $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$，其中 $d_i = H \times W \times C$ 对应图像中像素的高度 H，宽度 W 和通道数 C 的乘积。理想的聚类结果应该是每个图像簇包含有相似图像。这里相似图像的概念来自于图像分类数据集，指代图像内容包含有相同的目标物。图像特征编码可形式化表示为

$$\boldsymbol{z} = M(\boldsymbol{x}). \tag{4-1}$$

其中 M 表示图像特征提取模型，$\boldsymbol{z}$ 表示对图片 $\boldsymbol{x}$ 提取的特征编码。$\boldsymbol{z}$ 表示为 n 维列向量。

对高维图像进行特征编码主要存在两个问题。第一，高维图像数据在图像空间的分布是稀疏的。由于图像间的距离难以度量，直接地聚类图像是不合理的，需要提取特征编码。在深度图像聚类中，编码过程常用特征编码网络解决，但当图像样本过少时，需要搭建合适的网络结构，且训练学习紧密的特征表达变的困难。第二，在图像上定义图像语义距离是困难的。为了实现相似图像内容的聚类，常将图像映射到语义特征编码，并在语义编码空间进行聚类。最近，相关的研究主要围绕采用深度网络学习可区分性的语义特征编码展开。深度嵌入聚类 (Deep embedded clustering, DEC) [83] 利用卷积网络来学习图像的特征表示，并利用 Kullback-Leibler (KL) 散度损失迭代优化编码网络的聚类损失。IDEC[84] 方法结合自动编码器中的重建损失和聚类损失来改进 DEC 中学

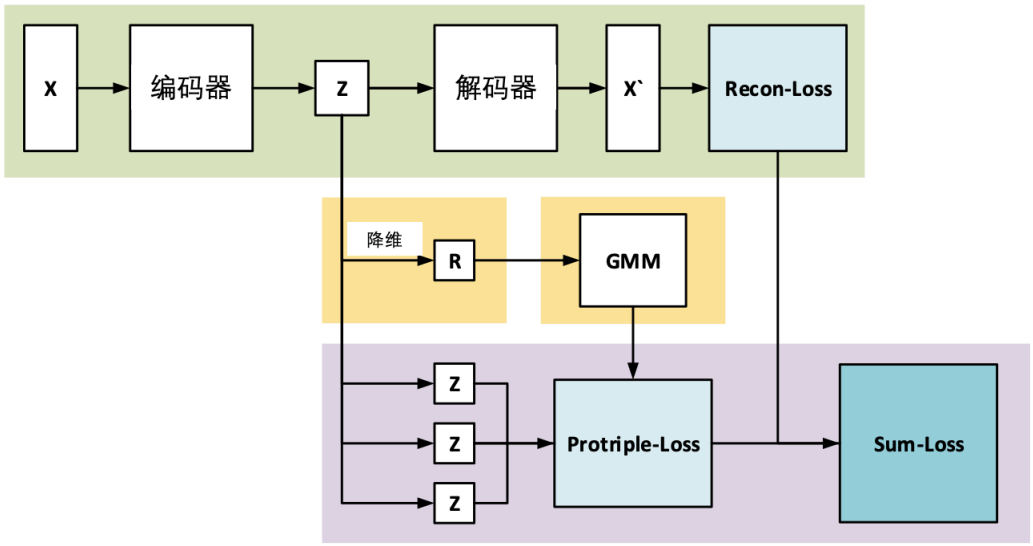


图 4-1: DERC 聚类模型结构示意图。

习的语义特征。DEC 和 IDEC 等方法都直接在编码器网络的语义空间中进行聚类，例如 DEC 采用 k-means 方法对特征编码进行聚类。深度正则嵌入聚类 (Deep embedded regularized clustering, DEPICT)[85] 对特征编码采用层次聚类方法。然而，由于编码器的学习过程和聚类过程的分离，这些图像聚类方法在学习编码网络的过程中无法确保所学习到的特征编码空间是符合下游选取的聚类方法。即无法明确特征编码空间适合于何种聚类方法，这限制了上述方法泛化性能。

4.2.2 DERC 模型

本节提出了深度图像特征编码方法（Deep Embedded dimensionality Reduction Clustering，DERC），并将其应用到图像聚类过程中，具体分为如下三个部分。首先，利用自编码网络模型，无监督地从待划分的图像数据集中学习特征编码。其次，对网络学习到的图像特征进行降维分析。最后，对降维后的图像特征进行聚类，并划分图像数据集。如图 4-1所示，DERC 聚类方法用于无监督地学习图像特征，并解决图像聚类任务。在图 4-1中，绿色部分为 DERC 方法的网络模型部分。模型部分采用自编码网络模型，其中编码器用于学习和提取图像特征，解码器用于对特征进行解码生成原始图像。黄色部分为对图像语义编码的初步聚类过程。由于编码器网络提取的图像特征仍旧为高维的特征向量，但选取高斯混合聚类模型适用于处理低维数据。因此，在进行聚类之

前，DERC 引入降维操作，将图像特征映射到适合聚类方法的低维空间。本节中将其称为图像语义特征的降维空间。基于对图像特征编码的降维可视化分析，DERC 采用无监督的高斯混合模型对降维后的图像语义编码进行聚类。紫色部分表示将高斯混合模型生成的聚类伪标签用于编码器网络的训练，学习更具有区分性的图像特征。DERC 提出了概率三元组损失，用于编码网络学习图像特征的过程。图 4-1 总体展示了 DERC 中各个组成模块的大体内容。本节的如下内容分别对各个模块进行详细描述。

自编码器网络：首先，DERC 模型主体采用自编码网络结构，由编码器和解码器两部分网络组成。编码器网络用于抽取图像的语义特征，完成对图像的特征编码，解码器对特征进行解码生成原始图像，用于辅助编码器的训练。设图像 $\boldsymbol{x}_i$ 经编码器网络非线性映射为图像编码 $\boldsymbol{z}_i \in \mathbb{R}^{d_z}$, 其中 d_z 为特征编码空间的维度且 $d_z << d_i$。编码器网络被看作为函数 $E_\theta : X \rightarrow Z$, 实现图像空间 X 到语义编码空间 Z 的映射。E_θ 代表编码器，其中待学习的参数为 θ。编码器网络由六至八层的卷积神经网络搭建而成，将二维平面上的的图像映射到 128 维的 $\boldsymbol{z}_i$。解码器网络作为函数 $D_w : Z \rightarrow X'$, 实现语义编码空间 Z 到生产图像空间 X' 的映射。解码器的网络结构与编码器类似采用上采样卷积层搭建。同时，为确保特征编码 $\boldsymbol{z}_i$ 对图像的表示能力，在编码器和解码器之间没有采用跳层连接的方式。

降维聚类过程：其次，对于图像的特征编码，DERC 进行了初步的聚类及可视化分析。图像数据集 X 通过学习后的编码器映射到特征编码 Z。由于图像特征编码也是高维数据，在选取合适的聚类方法时，会面临三个主要难点。第一，许多聚类方法需要提前确定聚类的个数。由于图像聚类的数据集借鉴图像分类任务，本文假定聚类个数与图像类别数一致。这种方法显然是取巧的，通常的做法是采取手工可视化分析辅助确定高维数据的聚类个数。第二，度量图像编码的距离相似度，用于聚类相似样本区分不同聚类簇。由于通过自监督学习的图像特征编码是高维数据，如何度量图像编码的距离识别相同目标，分离不同目标是需要进行分析和比较的。第三，编码空间中，图像特征向量存在稀疏性，数据分布不均衡。这显著影响了基于样本密度的聚类方法。为了解决上述难题，DREC 采用 t-SNE 方法对其进行了降维可视化分析。t-SNE 作为一种降维方法，其核心思想是在低维空间中重构高维空间中的数据分布。假设特征编码 $\boldsymbol{z}_i$，且以该特征点为中心的高斯分布的标准差为 $\boldsymbol{\delta}_i$ 的，则邻居编码节点 $\boldsymbol{z}_j$

到 $\boldsymbol{z}_i$ 的距离为：

$$w_{j|i} = \frac{\exp\left(-\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|\boldsymbol{z}_i - \boldsymbol{z}_k\right\|^2 / 2\sigma_i^2\right)}. \tag{4-2}$$

设 $\boldsymbol{r}_i$ 为 t-SNE 方法对 $\boldsymbol{z}_i$ 的低维嵌入表示，即 $\boldsymbol{r}_i \in \mathbb{R}_r^d$, d_r 为图像特征降维后的维度且 $d_r << d_z$。t-SNE 中假设空间 $\mathbb{R}$ 中的数据分布服从学生 t 分布。具体地说，邻居节点 $\boldsymbol{r}_j$ 到 $\boldsymbol{r}_i$ 的距离为

$$v_{j|i} = \frac{\left(1 + \left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \left\|\boldsymbol{r}_k - \boldsymbol{r}_l\right\|^2\right)^{-1}}. \tag{4-3}$$

该距离公式用于度量图像样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的相似度。t-SNE 采用梯度下降法迭代地减少特征编码分布与降维编码分布的 KL 散度，从而学习图像特征编码 $\boldsymbol{z}_i$ 的低维表示 $\boldsymbol{r}_i$。

对于降维后的二维或三维的编码向量集合，通过可视化分析，DERC 采用简单的高斯混合聚类模型对其进行初步的聚类。选择高斯混合聚类模型主要有三个原因，能够分别解决上述在图像特征空间聚类的困难。第一，可视化分析可以方便地确定聚类的个数，解决了高斯混合模型中需要提前指定聚类个数的问题。第二，在降维编码空间中，高斯混合模型可以采用欧式距离度量样本间的相似度。第三，从图像特征编码到降维表示后，图像数据可以集中表示，解决了高维空间中数据稀疏的问题。设含有 K 个分量的高斯分布混合模型 M，图像的降维编码 $\boldsymbol{r}_i$ 对该模型的采样概率为：

$$p_M(\boldsymbol{r_i}) = \sum_{j=1}^{K}\beta_j \cdot \mathcal{N}(\boldsymbol{r}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{4-4}$$

其中 $\mathcal{N}(r_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ 为第 j 个高斯分布量，β 表示第 j 个分布的概率且 $\sum_{j=1}^{K}\beta_j = 1$。则降维编码 $\boldsymbol{r}_i$ 来自于第 j 个分量的概率为：

$$p_{i,j} = \frac{\beta_j \cdot \mathcal{N}(\boldsymbol{r}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{p_M(\boldsymbol{r_i})}. \tag{4-5}$$

概率三元组损失：最后，编码器为了学习更具有区分性的图像特征，本文利用高斯混合模型的初步的聚类结果，提出了新的概率三元组损失函数，用于减少相似图像特征编码的类内距离，增加不同图像特征编码的类间距离。首

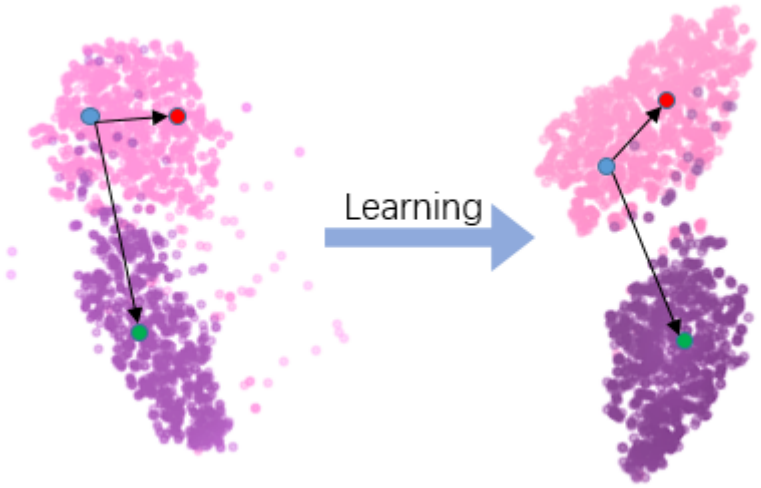


图 4-2: 可视化展示了概率三元组的学习过程。左半部分表示未学习前的图像特征分布，右半部分为利用概率三元组学习之后的图像特征分布。可视化结果来自于 MNIST 数据集中“3”和“5”图像的降维分析。

先，定义了图像特征编码间距离。设特征编码 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 间的欧式距离为：

$$d_{i,j} = \| \boldsymbol{z}_i - \boldsymbol{z}_j \|_2 . \tag{4-6}$$

我们试图利用降维特征的聚类结果学习图像特征 z，使其的编码距离分布符合高斯分布。我们选取由三个图像特征编码组成的采样集合，分别为锚点 $\boldsymbol{z}_i$，正样本点 $\boldsymbol{z}_p$，负样本点 $\boldsymbol{z}_n$。利用降维后的聚类结果，计算锚点所处的最大概率的聚类分量 k，并计算正样本和负样本处于聚类分量 k 的概率分别为 $p_{p,k}$ 和 $p_{n,k}$。概率三元组损失定义为：

$$\ell_c(z_i, z_p, z_n) = p_{p,k} d_{i,p} - p_{n,k} d_{i,n} + \text{margin}. \tag{4-7}$$

其中 margin 为预定义的正负样本到锚点的间隔距离，用于增加不同类别的特征编码区分度。图 4-2展示概率三元组对于图像特征编码的影响。 利用聚类生成的伪标签，编码网络学习了更具有区分性的图像编码，提升了聚类效果。

4.2.3 DERC 模型的训练过程

DERC 模型的学习过程分为两个阶段。在第一阶段的学习过程中，采用重构损失函数，通过自监督学习训练编码器和解码器的网络参数。具体地说，对

于输入图像 $\boldsymbol{x}_i$，编码器 E_θ 将其映射到对应图像语义编码 $\boldsymbol{z}_i$。解码器 D_w 利用特征编码 $\boldsymbol{z}_i$ 重构出原始的图像 $\boldsymbol{x}_i'$。$\boldsymbol{x}_i'$ 与 $\boldsymbol{x}_i$ 重构损失定义为图像像素间的 L_2 距离。编码器 E 和解码器 D 在数据集上的累计的重构误差为：

$$\mathcal{L}_{\mathrm{p}}(E,D)=\frac{1}{N}\sum_{i=1}^{N}\|D(E(x_i))-x_i\|_2. \tag{4-8}$$

利用式 (4-8) 的损失函数，采用梯度下降法学习网络参数 θ 和 w。

当上述损失稳定后第一学习阶段完成，编码器初步实现图像到语义特征的编码。将 N 张图像的特征编码表示为 $Z=[\boldsymbol{z}_1,..,\boldsymbol{z}_N]$，利用 t-SNE 方法进行降维获取对应的降维表示 $R=[\boldsymbol{r}_1,...,\boldsymbol{r}_N]$。DERC 通过可视化方法确定高斯混合模型的 K 个分量，并利用 EM 算法学习每个分高斯模型的隐参数 β 和分布参数，完成对降维后的特征 R 的初步聚类。利用重构损失函数，编码器可以学习部分图像语义的特征表示，但无法进一步学习更具有区分性的特征编码。为增加不同类别的图像特征的距离，编码器利用了提出的概率三元组损失，学习更加离散的特征编码。结合三元组采样，编码器 E 在训练数据集上的概率三元组误差为：

$$\mathcal{L}_{\mathrm{c}}(E)=\frac{1}{N}\sum_{i=1}^{N}\ell(\boldsymbol{z}_i,\boldsymbol{z}_p,\boldsymbol{z}_n). \tag{4-9}$$

其中 N 表示采样的规模。同时，引入重构误差可以有效避免避免编码器的退化，在第二阶段的学习中，自编码器在数据集上的总训练误差为：

$$\mathcal{L}_{total}(E,D)=\mathcal{L}_{\mathrm{p}}(E,D)+\alpha\mathcal{L}_c(E)F. \tag{4-10}$$

其中 α 为超参数，用于平衡两类损失。综上所述，算法 4.1总结了 DERC 模型及学习方法。

4.3 DERC 中图像编码的学习方法

DERC 提出的分阶段的训练方法，用于特征编码网络的无监督学习。本节从方法的训练过程角度出发，分析 DERC 训练方法中的不同的学习方法，并分析无监督图像编码器的学习过程。

算法 4.1 深度降维嵌入聚类方法 (DERC)

输入: 图像集 $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, 指定的聚类簇 K.
超参数: 分隔距离 λ, 超参数 α, 迭代轮数 $iter$, 三元组采样数 N.
输出: 图像的聚类分布 $C = [C_1, ..., C_n]$, 编码网络参数 θ.

第一阶段，利用重构损失函数 (4-8)，训练自编码器网络。
▷ 第二阶段训练编码器网络。
while 未到达迭代轮数 $iter$ **do**
  采用前向传播，利用编码器网络计算图像的特征编码向量: $Z = E_\theta(X)$.
  利用 t-SNE 降维方法，将图像特征编码向量映射到欧式空间:$R = t\text{-}SNE(Z)$;
  通过 EM 算法，在 R 上学习高斯混合聚类模型，完成图像特征聚类: $C, P = GMM(R)$;
  while 三元组采样轮数 N **do**
    构建概率三元组数据，并计算 $\mathcal{L}_c(\theta)$ (4-9);
    结合重构损失函数，计算总损失函数 (4-10)。
    采用梯度下降法，更新编码器网络参数 θ。
  end while
end while
返回: 图像的聚类结果 C, 编码器参数 θ.

4.3.1 DERC 中的自监督学习方法

从图像数据出发，将自编码网络看作一个端到端的学生模型 F，则 DERC 第一阶段的学习过程可以看作为自监督的学习过程。在该学习方法中，将图像 $\boldsymbol{x}$ 作为输入，编码器的输出可以表示为 $\boldsymbol{x}' = F(\boldsymbol{x})$。$\boldsymbol{x}'$ 为自编码器对输入 $\boldsymbol{x}$ 的重构图像。为了学习模型 F 的参数，我们采用在 L_2 距离下的图像的重构损失为

$$\ell_r(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|^2. \tag{4-11}$$

利用该重构损失，DERC 中的自编码器学生模型可以实现在无标注图像数据集上的自监督学习。然而，简单地将自编码器看作单一的学生模型是不利于进一步研究图像特征编码的。当模型完成自监督学习后，如果将其看作黑盒模型，对应的输入和输出对于研究图像特征编码而言是无意义的。即学生模型 F 与待研究的图像语义特征没有直接联系，需要深入学生模型的内部细节发掘特征编码。

4.3.2 DERC 中模型间的合作学习方法

从模型角度出发，DERC 的自编码网络由编码器和解码器两部分网络模型组成，表示为 $F(\cdot) = D(E(\cdot))$。编码器网络从图像提取特征编码为 $\boldsymbol{z} = E(\boldsymbol{x})$，解码器 Decoder 将编码特征映射到对应的相似原图像为 $\boldsymbol{x}' = D(\boldsymbol{z})$。将编码器和解码器分别看作学生模型，其中编码器网络用于从图像中学习语义特征。从图像中学习语义编码的过程可看作为编码网络的学习过程。因此，从学习方法的角度考虑，DERC 中在第一阶段的编码器的学习过程可以看作为编码器和解码器的合作学习过程，实现了图像语义特征的学习。从模型的角度出发，合作学习是解决无标注数据下的图像特征编码的一种有效方法。

根据式 (4-8), 编码器 E 作为学生模型实现图像到语义特征的映射，学习图像语义编码。解码器 D 对于编码器的输出编码进行解码，输出重构的图像。解码器 D 的损失函数定义为输出图像和输入编码的对应原始图像的重构损失。利用式 (4-11) 可完成对解码器网络的训练。对于编码器网络 E，当输入图像 $\boldsymbol{x}$，其编码器网络输出的图像编码 $\boldsymbol{z}$ 由解码器网络进行标注，产生监督信息实现编码器的学习。因此，将 DERC 模型看作编码器网络和解码器网络两个学生模型，DERC 第一阶段的学习过程可以被看作为两个学生模型间的合作学习。编码器网络被作为图像到语义特征的编码网络，该合作学习过程可被用于无监督的学习图像语义特征编码。

4.3.3 DERC 中的知识蒸馏方法

从模型角度出发，DERC 还引入了无监督的高斯混合聚类模型，用于划分降维后的图像特征。在 DERC 的第二阶段的学习过程中，利用参数学习后的聚类模型，可以为图像编码提供伪标签。将学习后的聚类模型看作为教师模型，则可以实现对编码器输出的特征编码的标注。通过式 (4-7) 中的概率三元组损失，聚类教师模型可以将自身知识通过标注特征编码的方式，指导给编码器学生模型。该指导学习是直接面向图像特征编码的，概率三元组实现相同图像类的特征编码分布符合降维后的特征在欧式距离度量下的分布。这解决了在第一阶段的合作学习无法控制编码器提取的特征编码的问题。

4.4 实验设计与分析

本章实验主要解决人脸图像的特征编码问题，并后续在此基础上实现细粒度的人脸编辑。目前，神经网络提取的人脸特征编码具有丰富的语义特征。由于高维编码特征的复杂性和多样性，确定高效的特征编码是后续进行人脸编辑的重要步骤。因此，本章利用提出的无标注图像的特征学习方法，解决任意人脸图像的编码问题。为了评估学习后的特征编码的优劣，我们将在手写和人脸两类数据集上验证特征编码的聚类性能。

4.4.1 数据集和评价指标

本章研究的图像聚类数据集，主要包括手写数字和人脸图像。对于手写数字图像，本章选取了 MNIST[86] 和 USPS [①]数据集。其中，MNIST 包含有 0 到 9 的手写数字共 1 万张图片，其大小为 32×32 的灰度图像。USPS 图像来自于从邮政收集的手写数字，大小为 16×16 的灰度图像。为研究 DERC 方法在人脸数据中的聚类效果，本章选取了 FRGC[②]，YTF[87] 和 CMU-PIE[88] 数据集。FRGC 数据集中的人脸图像主要集中在面部区域，其大小为 32×32 的 RGB 图像。YTF 的图像内容不仅包含人脸，还包括人的轮廓信息。该数据集由多组不同方位的摄像机对同一目标拍摄获取，包含有 41 类，其大小为 55×55 的 RGB 图像。CMU-PIE 数据集则包含有 68 个试验者，每个试验者分别做 4 组不同的人脸表情。表4-1总结了聚类实验所涉及的数据集的相关信息。

表 4-1: 图像聚类相关的手写数字和人脸数据集。

Dataset	# Sample	# Classes	Dimensions
MNIST-full	70,000	10	28×28×1
MNIST-test	10,000	10	28×28×1
USPS	11,000	10	16×16×1
FRGC	2,462	20	32×32×3
YTF	10,000	41	55×55×3
CMU-PIE	2,856	68	32×32×3

为了评估聚类的结果，分析学习的特征编码的优劣。本章采用聚类精度

[①]http://www.cs.nyu.edu/~roweis/data.html
[②]http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html

(Clustering Accuracy, Acc) 和归一化互信息 (Normalized Mutual Information, NMI) 作为评价指标。利用图像分类标签，通过匈牙利算法（Hungarian Algorithm）解决指派问题，即获取聚类簇群和真实图像标签之间的最佳匹配，将不同的聚类簇对应到图像类别。期望的图像聚类结果应该可以区分不同类别的图像。设集合 $Y = [y_1, y_2, ..., y_N]$ 表示 N 张图像的 K 个聚类簇的划分标签，集合 $C = [c_1, c_2, ..., c_N]$ 表示 N 张图片的 M 个类别标签。聚类精度用于评估每个图像聚类簇与对应的真实标签相同的划分准确率，定义为：

$$\mathrm{Acc}(Y, C) = \frac{1}{N} \sum_{y_i \in Y, c_i \in C} \mathbb{I}(y_i = g(c_i)). \tag{4-12}$$

其中，y_i 和 c_i 分别对应着图像 $\boldsymbol{x}_i$ 的聚类标签和类别标签，g 为聚类标签到类别标签的最佳匹配映射。归一化互信息用于评估图像的聚类分布和真实标签分布的差异，其定义为，

$$\mathrm{NMI}(Y, C) = \frac{2\mathbf{I}(Y; C)}{\mathbf{H}(Y) + \mathbf{H}(C)}. \tag{4-13}$$

其中 $\mathbf{H}$ 和 $\mathbf{I}$ 分别表示信息熵和互信息。即 $\mathbf{H}(Y) = -\sum_k P(y_i) log(P(y_i))$, $\mathbf{I}(Y; C) = \mathbf{H}(Y) - \mathbf{H}(Y|C)$。ACC 和 NMI 的取值范围在 $[0, 1]$ 之间，值越高代表聚类分布与类别分布越匹配，聚类性能越好。

4.4.2 对比和消融实验

对比实验：DERC 方法对比了其他聚类方法，包括 K-means，谱聚类 (Spectral Embedded Clustering，SEC）[84], 基于路径积分的凝聚聚类（Agglomerative Clustering via Path Integral Clustering，AC-PIC）[89]，深度嵌入聚类（Deep Embedded Clustering，DEC）[83], 深度正则嵌入表示聚类（Deep embedded regularized clustering, DEPICT）[85], 无监督联合学习（Joint Un-supervised LEarning, JULE）[32] 和判别式增强聚类（Discriminatively Boosted Clustering, DBC）[90] 等方法。同时，为了验证 DERC 中各模块的有效性，本节设置了消融模型 DAE 和 DERC-R。其中，DAE 代表 DERC 方法没有采用提出的概率三元组损失，缺少第二阶段的学习过程。DERC-R 只采用 DERC 中的自编码模型和高斯聚类，缺少对编码特征降维过程。

实验设置： DERC 的自编码器模型采用卷积神经网络搭建编码网络和解码网络。如表4-2所示，编码网络由六层卷积神经网络和一层全连接网络组成。经过编码器，图像 $\boldsymbol{x}$ 被映射到 d_z 维的图像特征向量 $\boldsymbol{z}$。在本节实验中，默认的

表 4-2: DERC 采用的编码器分层网络结构图。

网络层	参数
conv1	filters 3×3×2, stride 1×1, pad 0, RELU
conv2	filters 3×3×16, stride 2×2, pad 0, RELU
conv3	filters 3×3×32, stride 1×1, pad 0, RELU
conv4	filters 3×3×128, stride 2×2, pad 0, RELU
conv5	filters 3×3×256, stride 2×2, pad 0, RELU
conv6	filters 3×3×32, stride 1×1, pad 0, RELU
full7	filters 288×128

特征编码维度 d_z 为 128。解码器具有类似编码器的分层镜像结构，实现特征编码到图像还原。训练的过程采用 Adam 优化器，其中学习率为 0.001 和动量系数为 0.99。在训练的第一阶段，默认的最大迭代轮数为 100（训练整个数据集为一轮）。当 L_2 距离的像素损失呈现波动难以下降时，可以提前结束训练。采用 *t*-SNE 方法，对图像特征编码进行降维，并可视化研究降维后的数据，选择合适的聚类模型。本节选取高斯混合聚类模型，混合模型的分量 K 默认采用图像集中的类别数。在第二阶段学习时，结合概率三元组聚类损失优化编码网络，超参数 α 默认为 0.1，且优化器参数与第一阶段的相同。

聚类对比实验及分析：表 4-3和表 4-4分别验证了各类聚类算法在手写和人脸数据集上的性能。从表 4-3和表 4-4中可以看出，对于高维图像，无法只依赖图像间的 L_2 像素距离作为度量进行聚类。传统的 k-means，AC-PIC 等方法聚类方法无法胜任图像聚类任务，这些聚类方法需要利用特征编码才能实现图像聚类。深度神经网络用于提取图像特征，能够提升相同类别图像的聚类结果。深度聚类模型的性能差异主要体现在聚类方法的选取和模型的学习上。对比 DEC 方法，DERC 显示地利用高斯混合模型对编码进行聚类，解决了编码网络中聚类中心学习的波动。对比 DEPICT 方法，DERC 替换了层次聚类方法，加速了模型学习特征和聚类的速度。对比其他聚类方法，本文提出的 DERC 在聚类精度和归一化信息等指标上取得了最优结果。同时，在 YTF 数据集上，经过 *t*-SNE 方法降维可视化分析后，数据分布不太适用于高斯混合模型。我们选取基于密度聚类的 DBSCAN 方法 [91] 实现对特征编码的聚类。从对 YTF 数据集的聚类处理可以看出，DERC 方法可以根据不同的数据分布采用合适的聚类方法。因此，DERC 方法对于不同类型的数据具有良好的适应性。

表 4-3: 比较不同聚类方法在手写数据集上的聚类准确度（ACC）和归一化互信息（NMI）。其中，标记 (-) 代表无法获得有效的聚类结果。

Dataset	MNIST-full		MNIST-test		USPS	
	NMI	ACC	NMI	ACC	NMI	ACC
k-means	50.0	53.4	50.1	54.7	45.0	46.0
AC-PIC	1.7	11.5	85.3	92.0	84.0	85.5
SEC	77.9	80.4	79.0	81.5	51.1	54.4
DEC	81.6	84.4	82.7	85.9	58.6	61.9
JULE	91.1	96.1	91.1	96.0	91.0	95.0
DBC	91.7	96.4	—	—	74.3	72.4
DEPICT	91.7	96.5	91.5	96.3	92.7	96.4
DAE	78.9	82.4	77.5	80.5	75.8	72.4
DERC-R	91.4	95.6	90.7	94.9	92.3	96.2
DERC	**92.7**	**97.5**	**92.3**	**97.2**	**94.2**	**97.7**

表 4-4: 比较不同聚类方法在人脸数据集上的聚类准确度（ACC）和归一化互信息（NMI）。其中，标记 (-) 代表无法获得有效的聚类结果。标记（*）代表采用 DBSCAN 的聚类结果。

Dataset	FRGC		YTF		CMU-PIE	
	NMI	ACC	NMI	ACC	NMI	ACC
k-means	28.7	24.3	77.6	60.1	43.2	22.3
AC-PIC	41.5	32.0	69.7	47.2	90.2	79.7
DEC	50.5	37.8	44.6	37.1	92.4	80.1
JULE	57.4	46.1	84.8	**68.4**	**100.**	**100.**
DEPICT	61.0	47.0	80.2	62.1	97.4	88.3
DAE	54.5	40.9	—	—	73.8	48.1
DERC-R	65.2	49.1	90.2*	68.0*	94.3	83.7
DERC	**66.7**	**51.3**	**90.7***	65.8*	99.6	97.9

消融模型结果及分析：表4-3和表4-4中也展示了消融实验中设置的 DAE 和 DERC-R 模型的实验结果。对比 DAE 方法的结果，DERC 方法进行了第二阶段的学习。该结果侧面验证了第一阶段的合作学习，编码器无法学习有区分性的图像特征编码。同时，该对比结果也验证了指导学习在特征编码的重要性，

表 4-5: 在 MNIST 测试数据集上，关于不同超参数 α 在聚类指标 ACC 和 NMI 的结果。

α	0.1	0.3	0.5	0.7	0.9
ACC	96.60	96.76	96.76	98.25	96.71
NMI	91.85	92.12	92.22	93.10	92.06

以及提出的概率三元组的有效性。对比 DERC-R 和 DERC 的聚类精度 ACC 和 NMI，可以看出在第二阶段中进行聚类前，对编码特征的降维有利于提高高斯混合聚类模型的聚类精度。此外，从学习方法角度，消融实验中的 DAE 方法可以看作为编码器学生模型只采用自监督学习的过程。在此基础上，DERC-R 方法引入了聚类模型作为教师模型，用于指导特征编码网络（学生模型）的学习。如表4-3和表4-4中所示，DERC-R 引入独立学习后，提高了特征编码学生模型的性能。同时，对比 DERC-R 方法的结果，DERC 方法利用了降维方法，采用了更加有效的教师模型，提高了指导学习的效率。这从侧面反映了对同一个学生模型合理地采用多种学习方法进行训练，有助于提高该学生模型的性能。

4.4.3 超参数及可视化分析

超参数分析: 在第一阶段的学习中，编码器将图像映射到特征编码维度 d_z 是重要的参数。图4-3展示了不同特征编码维度在 DERC 及其消融模型上对图像聚类精度的影响。从图4-3可以看出，当图像的特征编码过小时（小于 60），该编码无法实现对图像特征的提取，难以确保聚类的精度。此外，本节默认的图像特征编码维度为 128 存在特征表示的冗余，DERC 中显示的降维过程缓解了冗余编码问题，使得聚类结果相对稳定（当编码维度大于 70）。在第二阶段的学习中，表4-5展示了超参数 α 对图像聚类精度的影响。超参数 α 用于平衡图像重构损失和概率三元组损失，且该参数较为鲁棒。

训练阶段特征聚类的可视化分析：面向网络模型的训练过程，本节研究不同学习阶段对图像语义编码的影响。图4-4可视化展示了不同训练阶段图像特征编码的分布。首先，随机初始化编码器参数，将图像映射到语义特征编码。从图4-4的最左侧一列可以看出，未学习的图像编码无法区分不同类别的图像。其次，自编码器经过第一轮的独立学习和合作学习后，编码器能够学习有代表性的语义特征。相同类别的图像在特征编码分布较为集中。然而，从第二列可以看出，学习的编码特征无法显著区分不同类别的图像，相似类别的图

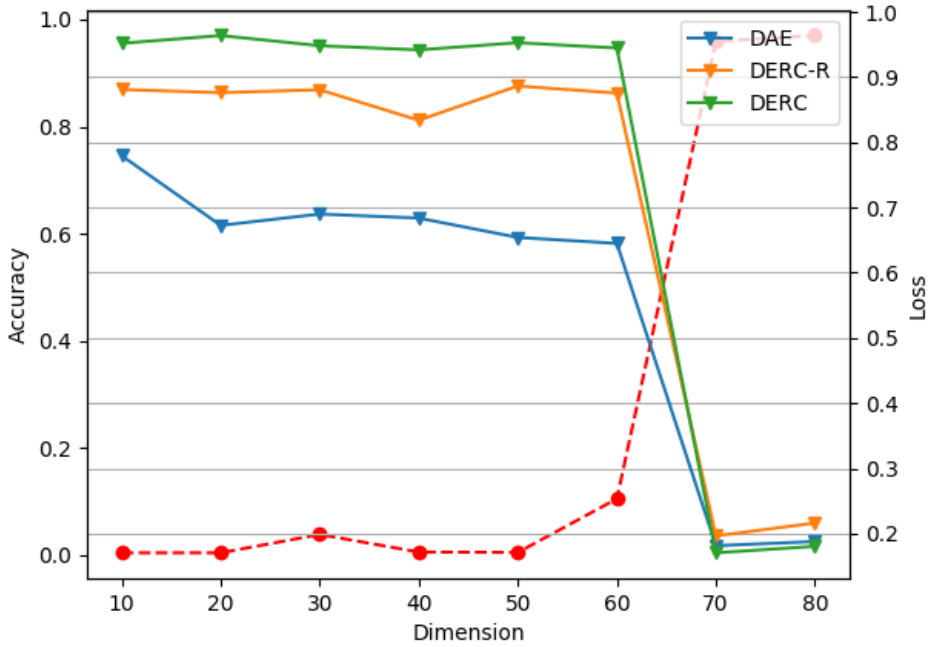


图 4-3: 不同特征编码维度对模型训练和聚类精度的影响。其中，红色虚线表示不同编码维度在 DERC 方法中的图像聚类精度。彩色的点实线表示采用不同训练方法后，不同编码维度的编码器的训练损失。

像的特征编码分布较为接近。最后，通过第二阶段的聚类模型生产伪标签用于编码器的指导学习，减少了图像特征编码的类内距离，增加了图像特征编码的类间距离。从第三列可以看出，学习后的图像特征编码分布较为离散，可区分性高。

特征编码及降维编码的分析：由于编码器提取的图像编码难以表示成低维数据，直接在特征编码 $\boldsymbol{z}$ 上进行聚类会面临一个主要的问题。即无法确定关于特征编码 $\boldsymbol{z}$ 的图像数据分布，难以选取合适的聚类算法。因此，DERC 处理降维后的特征编码 $\boldsymbol{r}$，并可视化分析后采用高斯混合聚类模型。图 4-5展示了不同特征编码下的聚类可视化结果。第一阶段训练结束后，利用编码器可以获取了图像的特征编码集合 Z。如图 4-5中第二列所示，当在图像编码集合 Z 上应用高斯混合模型进行聚类时，无法实现理想的聚类效果，部分类别存在划分错误。例如，在 USPS 数据集上，由于图像‘1’ 和‘7’的相似性，无法准确区分这两类的特征编码。在降维表示后的集合 Z 上，高斯混合模型对其有着良好的划分。对比在图像的特征编码集合 R 的聚类效果，高斯混合模型在降维表示集合 Z 能有效区分相互靠近的相似类别，学习过程也更加稳定快速。因此，在训

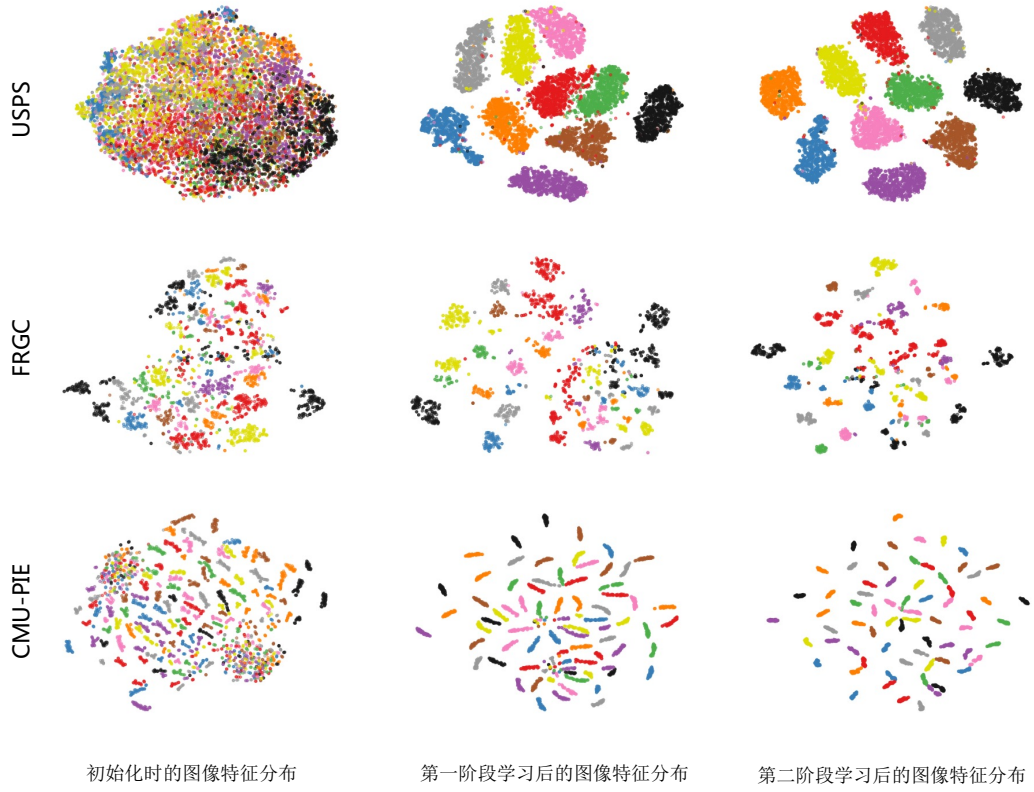


图 4-4: 可视化研究不同学习阶段学习到的图像特征。该图采用了 *t*-SNE 可视化方法，对编码的图像特征进行降维。其中，每个点代表单张图像，不同颜色代表不同类别的图像。在不同数据集上，进行特征空间的可视化分析。第一列的结果来自于随机初始化的特征编码器。第二列为第一轮学习阶段结束后的结果。第三列为编码器利用概率三元组损失进行第二轮学习后的结果。

练的第一阶段后，对图像特征编码进行降维后再聚类是必要的。在第二阶段的训练中，编码器利用概率三元组损失进行学习，提高了图像语义特征的表达。从式 (4-7) 可以看出学习后的图像特征编码的相似度符合欧式空间下的距离度量。在此图像特征编码上可以直接应用高斯混合模型进行聚类，并在聚类精度和归一化互信息等指标上，实现了先降维再聚类的类似精度。

人脸聚类可视化：在图 4-6中，DERC 方法对 FRGC 数据集进行聚类的样本结果。从 4-6可以看出，DERC 方法所提取的图像特征能够处理不同照明和轻微变形的人脸。在人脸聚类的问题中，DERC 也涉及了后文所研究的特征解码和编辑的部分研究内容。在 DERC 的自编码网络学习过程中，解码网路完成了特征到人脸的生成，可以看作为一种特征解码过程。但 DERC 的解码网络由于缺少有效的学习方法，其解码能力无法完成高质量的图像生成。并且，由于所提取的图像特征难以解释对应的语义内容，我们无法实现基于该特征编码的

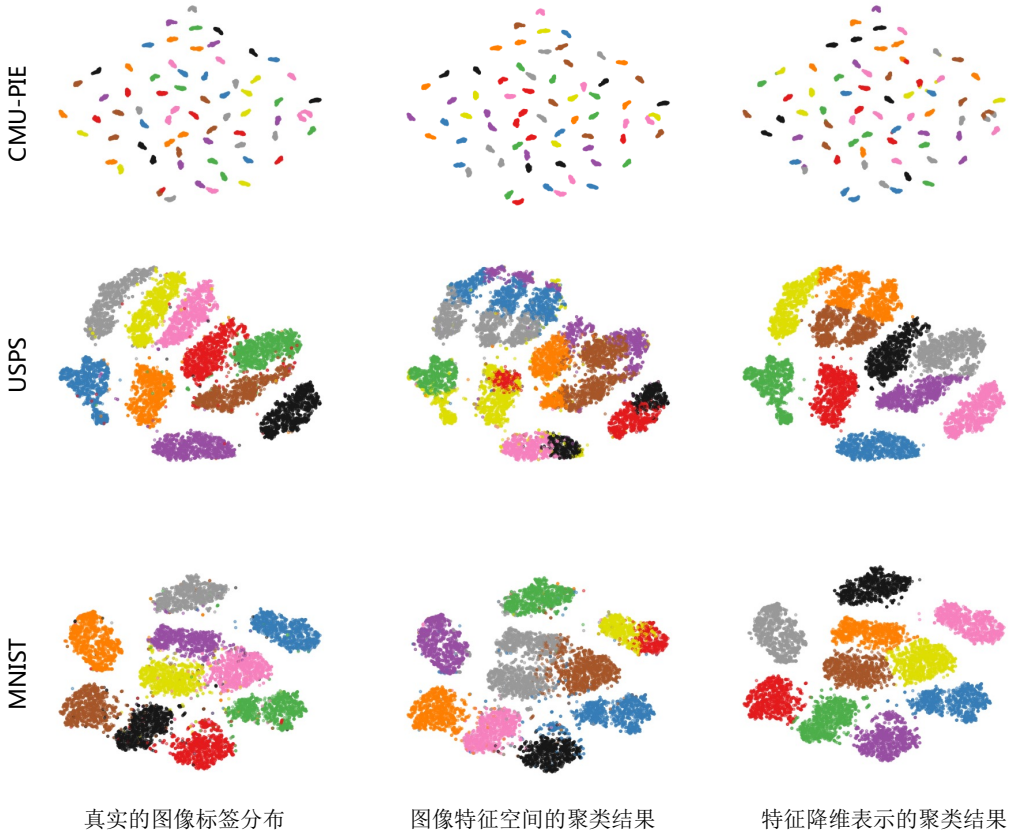


图 4-5: 在不同数据集上，可视化展示了利用高斯混合模型在特征编码和降维表示空间进行聚类的结果。该图采用了 *t*-SNE 方法对图像特征编码进行降维并可视化。

图像编辑。

4.5 本章小结

本章主要提出了无监督学习图像特征编码网络的问题，主要围绕如何学习编码器网络，实现对图像内容的语义特征编码展开。为了解决该问题，本章面向了具体的图像聚类任务，并提出了 DERC 聚类解决方法。DERC 方法实现了无标注图像的特征编码网络的学习问题，也在相关的图像聚类任务取得了良好的性能。同时，从学习方法角度分析了 DERC 方法，并对图像特征编码的学习方法进行了梳理。从数据上，DERC 方法采用自编码器模型完成图像到图像的自监督学习方法。从模型上，DERC 方法通过编码器和解码器进行合作学习，提取图像中的特征编码。利用知识蒸馏方法，DERC 方法将传统聚类方法作为教师模型，实现对编码器网络的指导学习。

图 4-6: DERC 方法在 FRGC 人脸数据集上聚类的可视化结果。其中，每一列都是同一聚类簇中随机选取的样本。

第五章　人脸编辑下的特征解码研究

5.1　图像特征解码及生成任务

根据下游的图像任务需求，图像语义编码可被用于分类和回归等网络模型。图像特征解码旨在研究在实际问题中图像特征的处理过程。对于编码器所提取的特征编码，依据下游任务选择合适的解码网络，实现图像特征到实际需求的解码。根据目标任务的不同，图像特征解码分别对应着分类、回归和生成等多种结构的网络模型。然而，这些研究更多关注在具体任务中的性能而忽略了图像特征解码的过程，难以深入研究图像特征在解码过程中的语义信息。为深入分析图像特征解码的过程，本章选择研究图像生成任务下的特征解码过程，大致有两个主要原因。其一，基于第4章中对图像特征编码的研究，图像生成任务下的特征解码的研究是图像特征编码的对偶问题。在上一章的图像语义编码中，所采用的解码器网络可以看作为一个具体的图像生成模型，实现了特征编码到图像内容的转换。其二，图像生成任务下的特征解码会生成易于理解的图像语义信息，具有良好的可解释性。特征提取和图像生成实现了对于图像的编码和解码的对应，可进行更深入地分析。

最近，图像生成的研究拓展了神经网络在图像上的的应用领域。深度学习下的图像生成主要利用网络模型生成各种伪造的逼真图像。第4章研究的自编码器的变种变分自编码器（Variational Auto-Encoders，VAE）[15] 是实现图像生成的一种通用模型。其中，隐变量编码 $\boldsymbol{z}$ 对应着图像特征编码。生成对抗网络（Generative Adversarial Network，GAN）借鉴对抗博弈的学习生成模型，解码随机高斯噪声并输出伪造的图像来拟合真实图像的分布。为了方便采样确保生成图像的多样性，变分自编码器和生成对抗网络都假设特征编码符合随机独立的高斯分布。然而，输入的随机噪声编码缺少可解释性，无法与图像特征编码建立联系。不同于以往关注生成图像质量的研究，本章重点将研究基于特征编码的图像生成任务，并解释图像特征在生成网络中的解码过程。首先，图像特

征可以来自于图像特征编码器对输入图像的提取，也可以由上述高斯噪声转换得到。从输入图像提取图像特征，再利用该特征实现特定的图像生成的研究，被归类为一大类图像翻译任务。图像翻译任务大都采用条件生成对抗网络，其在图像风格迁移、人脸替换和图像修复等任务中得到了广泛应用。依据上述不同的图像生成任务，本章聚焦于这些问题中所共同含有的核心内容，即基于图像特征的图像生成。在第4章的图像特征编码的相关研究的基础上，本章将关注图像特征的解码过程，并主要包括解码网络的架构及其学习方法。目前，采用编码器网络从图像中提取特征的方法被广泛研究，但将高斯噪声转换到图像特征编码的工作有待进一步的探索。StyleGAN[92] 模型推动了这一领域的研究。StyleGAN 提出将噪声编码映射到了图像的风格编码（Style 编码），并利用此编码实现高分辨率的图像生成。StyleGAN 中把 Style 编码看作控制生成内容的风格特征。不同于以上的观点，本章将 Style 编码看作为一种图像特征编码，并把 Style 编码的合成网络的图像生成过程被看作为图像特征解码过程。

本章将重点研究 StyleGAN 模型的特征解码过程及其学习方法。首先，分析了 StyleGAN 模型中的 Style 特征编码，并分析了其作为图像特征的合理性。受风格迁移的启发，StyleGAN 提出的 Style 特征本意是编码图像中的风格。在分析 StyleGAN 的解码过程中，我们发现仅将 Style 特征编码解释为图像的风格是难以解释图像的生成过程的。在合成网络的生成过程中，Style 编码还蕴含有丰富的语义信息，可控制图像内容的生成。其次，从学习方法角度，对 Style 特征编码的学习方法展开研究，梳理了合成网络（解码器）的学习过程。StyleGAN 网络中存在两种学习 Style 编码的方式，分别为生成器与判别器的生成对抗学习和利用预训练的图像特征编码器的知识蒸馏学习方法。其中，生成对抗学习主要学习 Style 编码在图像生成网络中的解码，知识蒸馏学习方法用于学习平滑的 Style 编码。最后，实验分析不同学习方法对于图像特征解码的影响。对比不同学习方法对 Style 编码的影响，验证了相关学习方法对训练图像特征编码的有效性。对于解码器学习后的 Style 编码，可以直接将其应用到图像分类中，并发现 Style 编码具有良好的可分离性。这为下一章中实现基于 Style 编码的特征可控编辑提供了相关的基础研究。本章的主要贡献有：

- 将 StyleGAN 生成模型中的 Style 编码看作为一种图像特征编码，并分析了基于 Style 特征解码的生成模型。对比基于噪声解码的生成模型，Style 编码有效提升了解码网络生成图像的质量。
- 研究图像生成模型的学习方法，分析了 StyleGAN 生成模型的学习方法，并

验证了生成对抗学习和知识蒸馏学习方法对训练 Style 编码的影响。

- 可视化分析了图像生成模型所学习的 Style 编码。将其直接应用到图像分类中，验证了 Style 编码在语义内容上具有可分离性，特征编码与语义内容存在关联。这些工作是下一章图像特征编辑的基础研究。

5.2 StyleGAN 上的图像特征解码模型

5.2.1 图像生成及 StyleGAN 模型

StyleGAN 实现了高分辨率的图像生成，是基于生成对抗网络生成图像的集大成者。图 5-1展示了 StyleGAN 的网络模型，其中左侧为 Style 编码的映射网络，右侧为生成图像的合成网络。原始的 StyleGAN 研究者将两个网络看作为整个 StyleGAN 模型，其被看作为简单的随机高斯采样变量到图像的生成过程，可形式化表示为

$$\boldsymbol{x}' = G(\boldsymbol{z}) \tag{5-1}$$

其中 $\boldsymbol{z}$ 表示由 n 个混合高斯分布采样的随机样本点，$\boldsymbol{x}'$ 为生成模型 G 所生成的图像。

为了研究 StyleGAN 的图像生成过程，我们将深入 StyleGAN 网络的内部结构，研究其中的 Style 特征编码。首先，利用映射网络（Mapping Network），将随机高斯分布噪声编码转换到 Style 编码。映射网络旨在实现对于随机噪声编码的解耦，并提高编码的可解释性。显然，独立同分布的高斯编码是不符合图像生成问题中生成的图像数据的分布的，且难以进行解释。在风格迁移的研究中，图像可以由内容和风格两部分控制生成。受风格迁移的启发，映射网络试图将其变换到 Style 编码空间并用于图像的伪造。映射网络由 8 层全连接神经网络组成。将高斯噪声 $\boldsymbol{z} \in \mathcal{R}^{512}$ 映射到 $\boldsymbol{w} \in \mathcal{R}^{512}$。

其次，合成网络（Synthesis Network）主要利用 Style 编码参数化控制指定内容的生成。合成网络采用卷积上采样的生成方式，由低分辨率逐层生成高分辨率图像。如图 5-1中右侧网络结构图所示，我们将合成网络中参数化的初始图像命名为图像模版（Image Template）。合成网络初始化了 512 个 4×4 分辨率的图像模版，用于表示不同的图像内容。在每层上采样的风格编码中，Style 编码 $\boldsymbol{w}$ 通过仿射变换映射为 A_i，并被应用到第 i 层的合成网络中，通过利用 AdaIN[93] 风格化控制生成图像的内容。设第 i 层的图像的内部生成图像为 $\boldsymbol{x}_i$，

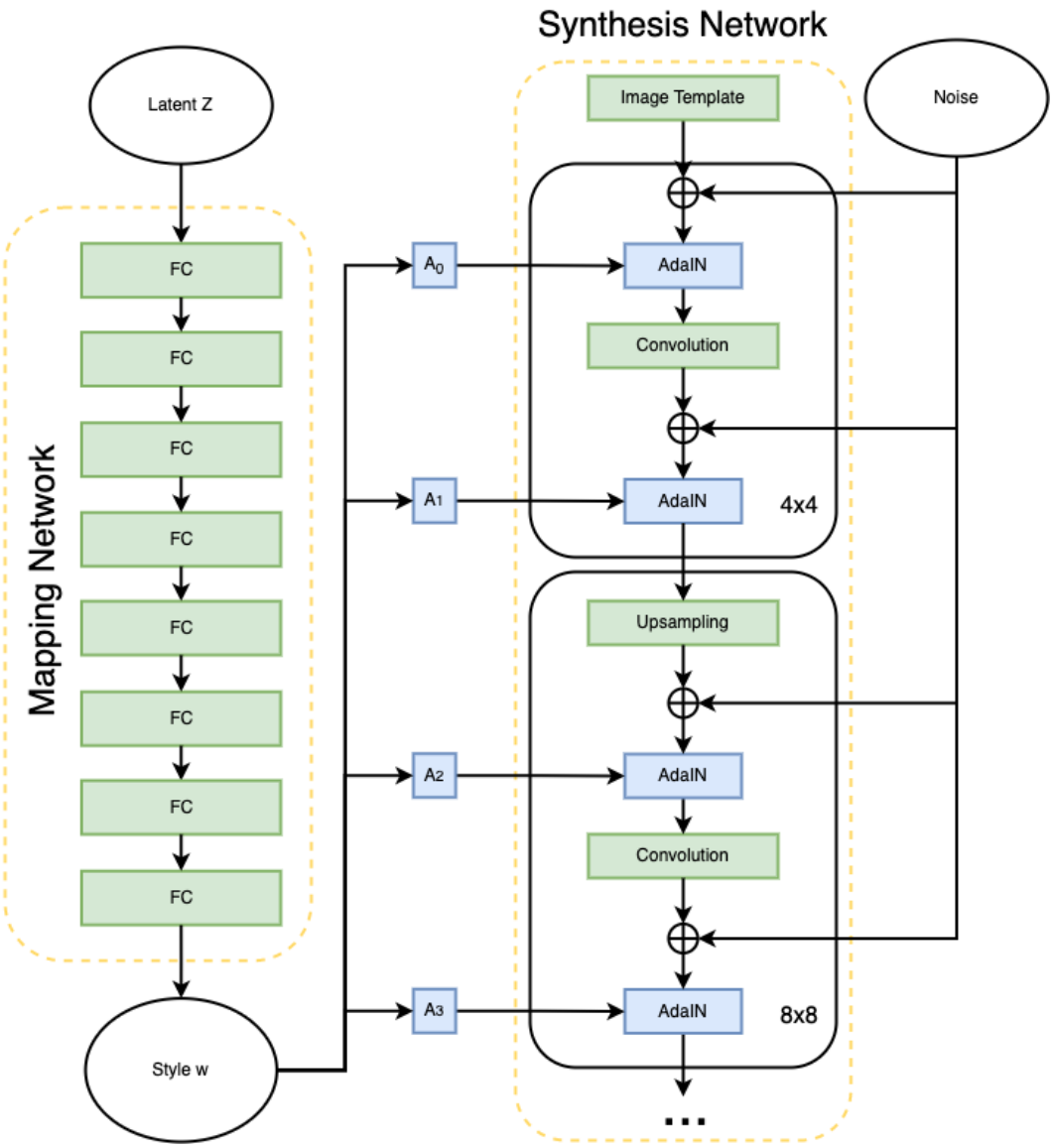


图 5-1: StyleGAN 网络模型结构示意图。其中，映射网络（Mapping Network）用于获取 Style 编码；合成网络（Synthesis Network）采用 Style 编码逐步生成伪造图像。

A_i 包含风格的均值和方差 $(\boldsymbol{a}_{m,i}, \boldsymbol{a}_{v,i})$。则第 i 层的 AdaIN 的风格迁移变换为：

$$AdaIN(\boldsymbol{x}_i, A_i) = \boldsymbol{a}_{m,i}\frac{(\boldsymbol{x}_i - \mu(\boldsymbol{x}_i)}{\delta(\boldsymbol{x}_i)} + \boldsymbol{a}_{v,i}. \tag{5-2}$$

其中，$\mu(\boldsymbol{x}_i)$ 和 $\delta(\boldsymbol{x}_i)$ 分别计算 $\boldsymbol{x}_i$ 的均值和方差。然而，归一化 $\boldsymbol{x}_i$ 会导致生成的内容产生人为的气泡瑕疵，影响图像生成的质量。在后续 Stylegan2[94] 的改进研究中删除对图像内容的归一化操作。本章后续也采用改进的风格迁移变

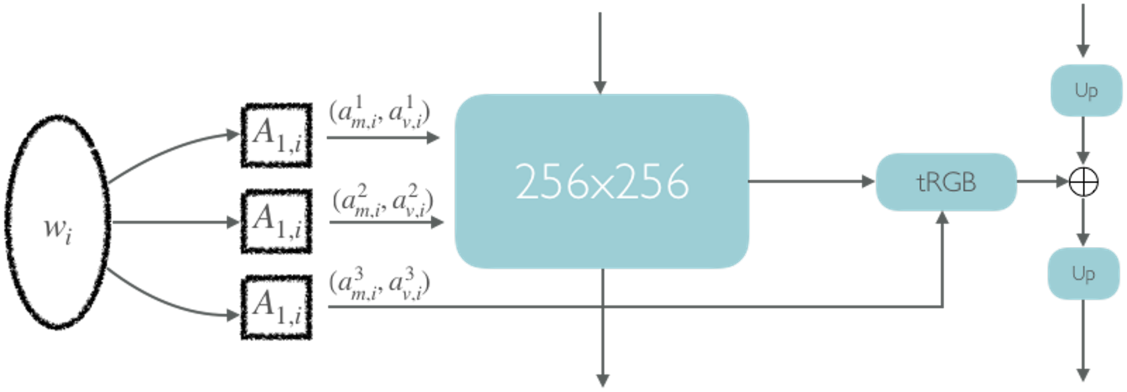


图 5-2: 在合成网络的上采样模块中，使用线性插值生成 RGB 图像的过程。

换为：

$$AdaIN(\boldsymbol{x}_i, A_i) = \boldsymbol{a}_{m,i}\boldsymbol{x}_i + \boldsymbol{a}_{v,i}. \tag{5-3}$$

此外，噪声（Noise）的引入是为了增强图像中细节部分的生成。

图 5-1实际上展示的是 Style 编码风格化图像内容的生成过程，其中省略了图像内容到 RGB 图像的生成过程。实际上，合成网络由 9 组的上采样模块组成。在每一层上采样模块中，利用 RGB 卷积网络层，完成图像内容特征到 RGB 图像的输出，并通过线性插值实现从 4×4 到 1024×1024 分辨率的图像生成。图 5-2详细表达了每个上采样模块中的内部结构，包含有图像内容激活层，Style 风格编码和 RGB 图像的生成。其中，w_i 由三组线性映射完成对三组卷积的风格化，其中第三组 $(a^3_{m,i}, a^3_{v,i})$ 用于 tRGB 层，完成图像内容特征图到 RGB 图像的风格生成。最右侧表示对图像的线性插值上采样，并完成由低分辨率到高分辨率的图像生成。在第 i 个上采样模块中，按照式 (5-3)，Style 编码通过两组独立的线性映射 $(a_{m,i}, a_{v,i})$ 实现对输入图像和上采样图像的风格控制。同时，每个分辨率图像也经由另一组 RGB 映射网络生成对应分辨率的 RGB 图像。最终生成的伪造图像由这些生成的不同分辨率的 RGN 图像上采样累加而成。

5.2.2 StyleGAN 模型的训练过程

StyleGAN 模型采用生成对抗网络的学习方法从独立同分布的高斯分布生成伪造图像，并拟合真实图像的分布。设 Style 网络模型为生成器 G，其中 G 网络由映射网络和基于 Style 编码的合成网络两部分组成。判别器网络 D 为衡量两个分布的 Wasserstein 距离，用于区分伪造图像和真实图像。G 和 D 网络模型

按照 WGAN-GP[95] 的方法进行训练。设特征编码 $\boldsymbol{z}$ 采样于随机高斯分布，对应生成的图像 $G(\boldsymbol{z})$ 为 $\tilde{x}$。$D(\tilde{\boldsymbol{x}})$ 表示为近似拟合的 Wasserstein 距离。对于从真实图像分布 $p_{data}()$ 中采样的图像 $\boldsymbol{x}$ 和从高斯混合分布 $p_z(\boldsymbol{z})$ 采样的 $\boldsymbol{z}$，生成模型 D 和判别模型 G 在此混合数据分布上的对抗损失为

$$\begin{aligned}\mathcal{L}_{gan}(G,D) =& \mathbb{E}_{x\sim p_{data}}D(x) - \mathbb{E}_{z\sim p_z}D(G(z)) + \lambda[\mathbb{E}_{x\sim p_{data}}(\| \nabla_x D(x)\|_2 - 1)^2 + \\ & \mathbb{E}_{\tilde{x}\sim p_z}(\| \nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2]\end{aligned}. \tag{5-4}$$

其中 $(\| \nabla_x D(x)\|_2 - 1)^2$ 和 $(\| \nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2$ 作为梯度惩罚项。它们的作用是让度量 Wasserstein 距离的判别器符合 Lipschitz 范数限制，即梯度的模尽可能的小于 1。StyleGAN 训练过程中，生成模型 G 和判别模型 D 分别最小化和最大化该损失函数，表示为

$$\min_G \max_D(\mathcal{L}_{gan}(G,D)). \tag{5-5}$$

当利用式 5-4的生成对抗损失训练生成器 G 时，我们将随机噪声编码 $\boldsymbol{z}$ 映射到隐空间中的 Style 编码 $\boldsymbol{w}$。然而，对抗损失函数无法用于指导 Style 编码的学习，难以学习解耦的语义特征。为进一步学习 Style 编码，增强生成图像的语义内容与 Style 编码的联系。StyleGAN 将生成器 G 拆分为的映射网络 F 和合成网络 S 两个部分。StyleGAN 提出了路径感知损失用于学习 Style 编码，使得在 Style 空间 W 上生成的图像内容较为平滑。设两个 Style 编码为 $\boldsymbol{w}_1 = F(\boldsymbol{z}_1)$ 和 $\boldsymbol{w}_2 = F(\boldsymbol{z}_2)$，经过合成网络 S 生成两张伪造图像 $S(\boldsymbol{w}_1)$ 和 $S(\boldsymbol{w}_2)$。为了学习 Style 编码的语义信息，StyleGAN 引入了预训练 VGG16 的图像特征编码器，来度量上述生成图像的相似性。路径感知损失定义了在 Style 空间中的两个 Style 编码的线形插值编码所生成的伪造图像。在图像特征编码器中语义特征中应该连贯平滑，具有局部相似性。StyleGAN 用预训练的图像特征解码器用于指导 Style 编码的学习。以步长 $\epsilon = 0.1$ 在 Style 编码 $\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 之间进行线性插值时（$t \in [0,1]$），在 W 空间中的两个 Style 编码的路径感知损失定义为

$$\mathcal{L}_w = E(\frac{1}{\epsilon^2}D_{VGG}(S(\text{slerp}(\boldsymbol{w}_1,\boldsymbol{w}_2;t), S(slerp(\boldsymbol{w}_1,\boldsymbol{w}_2;t+\epsilon))))). \tag{5-6}$$

其中，函数 $\text{slerp}(w_1,w_2;t)$ 用于计算编码向量间的球面线性插值（详情见附录 A.2），模型函数 D_{VGG} 采用预训练的 16 层的 VGG 图像分类网络度量了两张输入图像在图像特征编码上的欧式距离。

最后，StyleGAN 模型的训练过程混合采用 $\mathcal{L}_w$ 和 $\mathcal{L}_{gan}$ 损失函数，其中 $\mathcal{L}_{gan}$

损失主要负责解码 Style 特征并用于图像生成，$\mathcal{L}_w$ 损失用于学习映射网络生成的 Style 特征编码。

$$\mathcal{L}_{total} = \mathcal{L}_{gan} + \alpha \mathcal{L}_w. \tag{5-7}$$

其中，超参数 α 用于平衡上述两个损失的权重占比。

5.2.3 Style 编码及学习方法

StyleGAN 采用生成对抗的学习方法，训练图像特征（Style 编码）的解码器网络，用于生成逼真的伪造图像。从数据角度看，StyleGAN 只需收集感兴趣的图像数据，无需人为标记。模型的训练可看作为对整个数据集的拟合过程。从模型角度看，StyleGAN 不仅采用了生成对抗网络学习方法，还利用预训练的分类器特征构建感知损失，用于学习线性连续的 Style 编码。因此，我们主要分析来自于模型的标注信息，讨论 StyleGAN 中 Style 特征编码的合作学习和指导学习方法。

StyleGAN 中的生成对抗学习方法：从模型角度出发，StyleGAN 模型可以看作为整个学生模型 G，用于从噪声分布拟合生成图像分布。当 StyleGAN 采用 WGAN-GP 损失与判别器进行对抗学习时，可以将判别器网络也看作为学生模型 D，学习度量生成图像和真实图像之间的分布差异。对于判别器 D 而言，利用学生模型 G 生成的伪造图像和真实图像作为带标签的数据，并最大化式 (5-4) 中的 $\mathcal{L}_w(G, D)$ 对抗损失进行学习。对于生成器 G 而言，利用 D 标注生成的图像数据，最小化 $\mathcal{L}_w(G, D)$ 对抗损失进行学习。判别器 D 判别质量与生成模型 G 伪造图像质量相互影响。当判别器的分类能力过强时，生成器的图像分布与真实图像的分布重叠区域过小，会造成学生模型 G 难以利用该标注信息，沿合适的优化方向生成接近真实图像的分布。同时，生成模型 G 可能学习对任意的 Style 编码都生成一样的图像，造成生成模式的塌陷。然而，这种互相对抗的学习过程使得训练过程不稳定，式 (5-4) 难以收敛，生成网络的学习可能会陷入模式塌陷。

StyleGAN 中的知识蒸馏方法：在 StyleGAN 中的合作学习中，我们将整个生成模型看作为单一的学生模型，用高斯随机分布噪声去拟合真实的图像分布。Style 特征编码是隐含在生成模型中的，不便于本文对特征编码解码过程的研究。因此，基于学生教师模型的知识蒸馏学习方法，我们试图深入研究 Style 特征的编码和解码。从模型角度出发，StyleGAN 生成模型 G 可以由映射网络

M 和合成网络 S 组成。映射网络 M 利用随机高斯噪声生成 Style 编码，合成网络 S 通过 Style 编码控制不同图像的生成。在上节模型间的生成对抗学习中，我们未对 Style 编码进行学习，这使得其难以作为一种的图像特征。StyleGAN 通过引入了知识蒸馏学习方法，来学习 Style 编码的语义特征。具体的过程如下。首先，将预训练的图像分类 VGG 网络作为教师模型，通过式 (5-6) 标注合成网络生成图像，并生成 Style 编码的梯度信息。其次，将映射网络 M 看作为学生模型，利用由 $\mathcal{L}_w$ 感知损失生成的 Style 编码的梯度信息，优化该学生模型。通过上述的指导学习过程，StyleGAN 能够学习鲁棒的富含语义信息的 Style 编码特征。

5.2.4 图像特征编码与 Style 编码的关联

从上一小节中的 StyleGAN 模型的训练过程可以看出，Style 编码是隐式地嵌入在生成网络 G 学习过程中的。如果将生成模型看作一个整体，随机噪声编码 $\boldsymbol{z}$，难以与图像特征编码产生联系。因此，我们需要将生成模型 G 分离，看作为映射网络模型 M 和合成网络模型 S 的两个学生模型的组合。本小节中，我们研究映射网络模型编码的 Style 编码，并探究其是否可以作为一种图像语义编码。给定预训练的 StyleGAN 模型，生成随机采样点 $\boldsymbol{z}$, 通过映射网络 M 生成的 Style 编码 $\boldsymbol{w}$，并利用合成网络会生成对应的图像（噪声数据只会影响图像生成的部分细节内容）。Style 编码与生成图像内容呈现一一对应关系。根据第4章的图像编码研究，图像与特征编码也存在一一对应关系。本小节中，我们试图研究映射网络模型编码的 Style 编码，并阐明其可以作为一种图像语义编码，并可被用于多种下游图像任务。

为深入研究 Style 编码与图像特征编码的关系，我们希望将 StyleGAN 生成的图像重新映射回 Style 编码，并将其作为一种图像特征加以研究。首先，需要将图像重新映射回 Style 编码。获取指定图像的潜在 Style 编码大致有两种方式。其一，对于预训练的 StyleGAN 模型，固定网络内部的参数。对于指定单张图像，随机初始化的 Style 编码 $\boldsymbol{w}$，将其看作待优化的变量，利用反向传播的梯度下降法迭代优化 $\boldsymbol{w}$ 使其生成图像逐步拟合指定的图像。这种迭代地优化方法能对任意图像生成对应的 Style 编码，但生成效率较低且易受编码随机初始化的影响。其二，采用类似在第4节中的图像编码器，学习图像到 Style 编码的映射。具体来说，利用预训练的 StyleGAN 模型随机生成图像和 Style 编码对，用于训练特征编码器。Pixel2Style2Pixel[64] 采用此方法成功实现了对指定图

像的 Style 编码抽取，并拓展基于预训练的 StyleGAN 模型的研究。PixeltoPixel 不仅训练了真实图像到 Style 编码的图像编码网络，还训练了多种图像任务场景下的 Style 编码器。例如，在人脸的生成上，Pixel2Style2Pixel 成功训练了语义掩码到 Style 编码的编码网络，通过编辑语义掩码控制生成人脸。该模型还尝试从低分辨率人脸中学习 Style 编码，并直接利用 StyleGAN 的合成网络完成对高分辨率人脸的重建。基于 Pixel2Style2Pixel 的工作，Encoder4Editing[64] 提出了增量式的 Style 编码网络，改进了图像到 Style 编码的编码网络结构。StyleCLIP[57] 利用 StyleGAN 模型和 CLIP 模型，将文本编码和 Style 编码进行对齐，实现了通过自然语言对图像的生成对应的 Style 编码，并完成图像的生成。上述关于图像生成的研究工作都基于 Style 编码展开，并成功将其作为一种表示图像内容的特征编码。

5.3 实验与分析

本章主要验证了利用 Style 特征编码进行人脸生成的解码过程。我们研究特征的解码过程，并探究高维特征编码与人脸图像的对应关系。首先，对比其他生成模型，展示了 StyleGAN 模型在生成图像上的性能。可视化分析 Style 编码在合成网络中的解码过程。其次，从模型的训练角度出发，分析了不同学习方法下图像特征解码的影响。最后，分析了生成模型学习后的 Style 编码，并将其作为一种图像特征用于相关问题。

5.3.1 数据集和评价指标

本章研究的图像生成主要面向人脸头像数据集。人脸数据集主要包括 FFHQ（Flickr-Faces-HQ Dataset）[92] 和 CELEBA（CelebFaces Attributes Dataset）[96] 数据集。

FFHQ 数据集: 该数据集包含有 70000 张分辨率为 1024×1024 的高清人脸头像。该数据集具有众多人脸属性，拥有不同的年龄、性别、种族、肤色、表情、脸型、发色及姿态等，并囊扩各类物体配件，如眼镜、帽子、耳环等。

CELEBA 数据集: 该数据集包含超过 20 万张具有不同分辨率的名人头像。每张图像都进行了 40 种属性类别的标注。图像的内容涵盖了较大的人脸姿势变化和不同的背景。

本章采用弗雷歇感知距离（Fréchet Inception Distances，FID）[97] 用于评

表 5-1: 对比基于 Style 编码与噪声编码的生成模型在图像生成质量 (FID) 差异。

数据集	FFHQ	CELEBA	LSUN
GAN	8.04	7.76	6.57
StyleGAN	4.40	5.06	3.75

价生成图像质量。FID 将生成图像的分布与一组真实图像的分布进行比较，度量这两个分布的差异。首先，对于图像而言，FID 采用预训练的 Inception 图像特征提取模型提取图像的特征。其次，对于生成模型伪造的图像数据集 $\mathcal{I}_g$ 和真实的图像数据集 $\mathcal{I}_r$，利用特征编码网络提取对应的图像特征集 $\mathcal{F}_g$ 和 $\mathcal{F}_r$，并计算特征集向量的均值和协方差矩阵。设生成的伪造图像集的特征向量的均值和协方差为 $(\boldsymbol{u}_g, \boldsymbol{\Sigma}_g)$，真实图像集的特征向量的均值和协方差为 $(\boldsymbol{u}_r, \boldsymbol{\Sigma}_r)$。最终，FID 度量了在生成图像集和真实图像集上的特征向量上的距离为

$$FID(\mathcal{I}_g, \mathcal{I}_r) = ||\boldsymbol{u}_r - \boldsymbol{u}_g||^2 + Tr(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_r)^{1/2}). \tag{5-8}$$

其中 Tr 为矩阵的迹。在下列实验中，我们选取 50000 张真实图片和生成 50000 张伪造图片用于评估 FID 指标。

5.3.2 实验设置

所有的实验均在服务器上运行。该服务器拥有 40 核 cpu 和 4 张 1080TiGPU，含 128G 运行内存。在预训练 StyleGAN 的模型中，StyleGAN 的生成器模型 G 采用 Adam 优化器更新参数，学习率为 0.002，动量系数为 $(0, 0.99)$。判别器模型 D 也采用 Adam 优化器，学习率为 0.002，动量系数为 $(0, 0.99)$。StyleGAN 在 4 张 GPU 上并行训练，BatchSize 的大小为 32，迭代 25000 轮。在式 (5-7) 的总损失函数中，超参数 α 设为 0.5。

5.3.3 Style 编码解码过程的实验及分析

首先，验证了引入 Style 编码对图像生成质量的影响。其次，分析了合成网络中不同层中内容的生成。

对比 Style 编码和噪声编码的图像生成质量：为了验证 Style 编码对生成图像质量的影响，我们对比了传统的 GAN 模型和 StyleGAN 模型在生成图像质量

图 5-3: 可视化基于 Style 编码生成的人脸头像。

上的 FID 指标, 其中 GAN 生成模型直接实现随机噪声到图像的映射。GAN 模型删除了 StyleGAN 中的映射网络 M，只利用合成网络实现对高斯噪声解码。为了确保模型对比的公平性，在训练时，我们只利用式 (5-4) 中的对抗损失进行训练，并且其余的超参数保持一致。表 5-1展示了 StyleGAN 和 GAN 模型在人脸和动物图像上的 FID 指标。如表 5-1所示, StyeleGAN 网络提出映射网络 M 将随机噪声映射到 Style 编码，并对 Style 编码进行解码实现图像的生成。基于 Style 编码的合成网络显著提升生成图像的质量。

Style 特征编码的生成图像可视化分析：图 5-3展示了基于 Style 特征编码生成的人脸头像。虽然 StyleGAN 模型能够生成逼真的高清人脸图像，但在生成整体稳定性上仍存在不足。如图 5-3中所示，每张人脸中仍存在肉眼可辨识的瑕疵。这种生成的瑕疵在 StyleGAN 生成模型中普遍存在，消除这些瑕疵能有效提高图像的质量。因此，在下一章的图像特征编辑中，我们试图研究 StyleGAN 的内部隐空间及其语义的生成过程，利用内部特征编码实现指定图像内容的编辑。

分析 Style 编码在合成网络中的分层生成效果：在 StyleGAN 模型中，通过多个线性映射网络，Style 编码被分别应用到合成网络的每个生成层中。为了分析图像由从低分辨率 4×4 到 1024×1024 高分辨率的生成过程，将 Style 编码应用于低分辨率层并将高分辨率层的 Style 编码置为零。这种操作的目的在于控制 Style 编码在合成网络中的表达，以可视化内部图像生成层。如图 5-4所示，在合成网络的前面低分辨率的生成图像中，Style 编码用于控制生成人脸的轮廓

及姿态。在中间的内容生成层中，Style 编码用于人脸局部物体及细节的生成。在最后面生成层中，Style 编码用于整体图像的亮度调节。Style 编码网络在合成网络进行分层表示，不同层的编码控制着不同的语义内容。因此，在某种程度上，Style 编码存在语义内容上的耦合，无法明确通道中所代表的图像特征。

表 5-2: 基于 Style 编码的不同人脸属性的分类精度。

分类属性	微笑	口红	老年	胡子	卷发
ACC（%）	90.4	97.3	88.2	95.4	80.6

将 Style 编码作为特征用于其余图像任务： InterfaceGAN 方法 [27] 直接将 Style 编码作为一种图像特征应用于人脸属性分类。借鉴上述的研究思路，在 CelebA[96] 数据集上，我们训练了基于 Style 编码生成图像的多种人脸属性二分类器，并计算利用 Style 编码进行分类的精确度。表 5-2展示利用 Style 编码对多种同人脸属性进行分类的准确度。如表 5-2所示，Style 编码在多种人脸属性的表示中具有较强的可区分性。这为下一章研究特征编码与潜在语义内容，并提出基于编码的图像内容编辑方法做了相关铺垫。

合作学习和指导学习的对比研究：本节利用相关的深度学习方法分析 Style 特征编码的学习过程。首先，根据式 5-4中的对抗损失函数, StyleGAN 的生成模型 G 和判别模型 D 采用生成对抗学习方法进行训练。为了进一步学习 Style 编码，我们将生成模型 G 分解为映射网络和合成网络，利用式 (5-6)，采用知识蒸馏方法对合成网络 M 进行指导学习。生成对抗学习方法用于隐式的学习 Style 编码，知识蒸馏学习方法用于显式地学习 Style 编码。这两种学习方式可以分开单独训练，也可先进行生成对抗学习再用知识蒸馏方法进行微调。另外，根据式 5-7, 采用混合训练的学习的方式训练 StyleGAN 模型。如表 5-3和表 5-4所示，比较了采用不同学习方法对训练结果的影响。从表 5-3可以看出，生成对抗学习主要用于训练 StyleGAN 模型中的合成网络，并且知识蒸馏方法无法显著提升图像生成的质量。然而，表 5-4展示了生成对抗学习能有效提升 Style 编码的质量。知识蒸馏方法用于学习紧凑的 Style 特征表达，减少了感知路径损失。

表 5-3: 比较不同学习方法对于 StyleGAN 模型的生成图像质量（FID）的影响。

学习方式	生成对抗学习方法	知识蒸馏方法（微调）	混合学习方法
FFHQ	5.04	5.37	4.40
LSUN	3.88	4.13	3.75

表 5-4: 比较不同学习方法对于 Style 编码的感知路径损失。

学习方式	生成对抗学习方法	知识蒸馏方法（微调）	混合学习方法
FFHQ	223.5	153.4	122.5
LSUN	1484.6	1006.3	981.6

5.4 本章小结

本章主要面向基于 Style 编码的图像生成问题，并围绕预训练的 StyleGAN 模型展开研究。我们选择与人脸生成相关的预训练的 StyleGAN 模型，并深入研究和分析其中 Style 编码的解码和学习过程。首先，阐明 Style 编码属于一种图像特征编码。对比随机高斯噪声编码，Style 编码有效提高了图像生成的质量。其次，梳理了 Style 编码的学习方法。我们将 StyleGAN 模型分解为生成模型 G 和判别器模型 D，并采用对抗损失进行两个学生模型的生成对抗学习。同时，将生成模型 G 分解为映射网络 M 和合成网络 S。StyleGAN 把预训练的 VGG 图像特征编码网络作为教师模型，指导映射网络模型 M 的学习，属于一种知识蒸馏学习方法。最后，实验验证了 Style 编码在图像解码过程中的优越性，也分析了不同学习方法对 Style 编码和解码性能的影响。

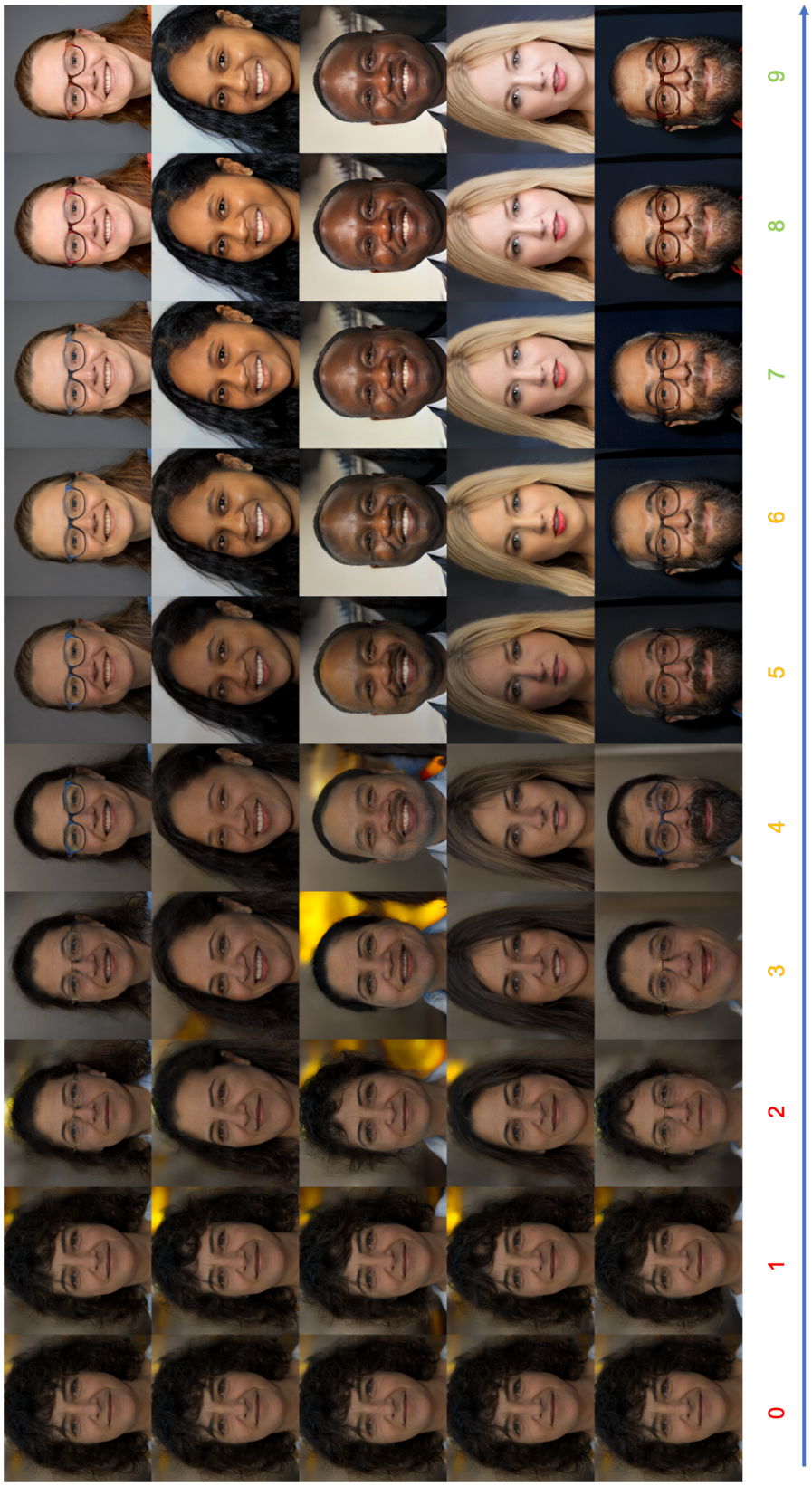


图 5-4: 在 FFHQ 人脸数据集上，分析 Style 编码在合成网络中的逐层解码的生成结果。

第六章　可指定属性的人脸编辑研究

6.1　基于图像特征的人脸编辑研究

基于上述图像特征编码和解码的研究，运用相关学习方法训练不同功能的神经网络模型，用于实现图像空间与特征空间的双向映射。在第4章图像编码的研究中，提出了利用无标注数据学习图像特征编码的方法，并利用该特征编码解决了图像聚类问题。在第5章图像 StyleGAN 生成模型的研究中，我们着重分析了 Style 编码与图像特征编码的联系，并初步验证了该编码与语义内容的关联。本章试图理解并解释图像特征编码与图像语义内容的对应关系。

要实现通过修改特征编码编辑图像内容，分析特征编码与生成内容的关联是其核心步骤。图像中包含有不同物体，复杂空间结构等抽象的语义内容，这种信息对于人类容易感知，但在计算机中却难以表示。为了在计算机中处理的这些图像信息，深度学习通常采用编码网络将图像映射到特征编码空间，将原始图像中的丰富信息表示为特征编码。然而，对于高维的特征编码向量，人类却无法直观理解特征编码所代表的语义信息。为理解图像特征编码实现可控编辑，本章结合前述两章的研究内容，将深入分析图像特征空间。研究图像特征空间的重点在于理解特征编码的语义含义，解释高维特征向量中部分分量所表示的图像内容。当探明特征编码与图像内容的关联后，本章还提出编辑图像特征的方法以实现细粒度地控制内容生成。即通过编辑图像编码实现可控的图像生成，完成图像编辑任务。

图像编辑任务旨在实现对图像的可控修改。对于神经网络模型而言，图像编辑是通过特定的方法控制生成模型所输出的图像内容。目前，基于神经网络的图像编辑研究主要分为两大类方法。第一类方法，通过控制网络模型的输入，来实现特定图像内容的编辑。该类方法大都采用条件生成对抗网络模型，并需要构造属性分类数据集，以此训练图像编辑模型。第二类方法，试图发掘预训练生成模型中的特征编码的语义信息，在隐特征空间中解析控制不同属性

的特征编码，并对其进行编辑完成对指定图像内容的修改。基于特征编码的图像编辑通常需要解决三个难点问题。其一，需要确定具体研究的生成模型和隐特征编码空间。由于解码器大都采用多层堆叠网络模块的设计，神经网络模型中内部存在多种隐编码空间。不同编码空间下所关联的语义内容存在不同程度的耦合。选取合适的特征编码空间进行分析是至关重要的。同时，由于随机训练的影响，同一个结构的模型即使在生成的图像上差距微小，但内部的隐特征空间的结构也会存在较大的差异。在进行编码语义分析前，得明确预训练的图像生成模型。其二，需要分析特征编码与图像内容的潜在映射关系，并定位相关特征通道所编辑的语义信息。一种简单但有效的方式是通过随机修改特征编码并观察生成图像上的改变，并选择有效的编辑向量对其加以解释，以实现对特定生成内容的编辑。这种扰动方法不仅探明了特征编码与语义内容的关联，还完成了特征编码的编辑并部分实现了图像的可控编辑。例如，GANSpace[40] 采用 PCA 方法分析了 StyleGAN 中的 Style 编码空间，将主成分特征向量用于对特征编码的扰动，并查明对应生成图像的编辑内容。然而，这种扰动式的方法需要手工确定不同特征编辑方向所对应的语义属性，无法针对指定的语义属性找寻对应的特征编辑向量。其三，确定不同属性特征编码可编辑的范围。当探明指定属性对应的隐特征编码通道后，即可实现对编码通道的编辑。为了确保编辑的强度符合模型的输入范围以生成合理的图像，需要确定编码方向上的的编辑强度。例如，StyleSpace[62] 利用指定属性的编码的峰值信噪比确定了单编码通道的编辑范围。

由于本章仍旧关注特征编码在图像编辑上的研究，我们选择第二类解决方案处理图像编辑任务。设图像特征编码为 $\boldsymbol{s}$，并由预训练的 StyleGAN 模型生成原始图像 *boldsymbolx*。基于特征编码的人脸属性编辑可形式化定义为

$$\boldsymbol{x}' = G(Edit(\boldsymbol{s})) \tag{6-1}$$

其中 $Edit()$ 为本文研究的人脸图像特征编辑方法，其作用于特征编码 $\boldsymbol{s}$。$\boldsymbol{x}'$ 为指定编辑属性内容后的生成人脸图像。

针对上述的编辑问题，本章分别给出了合适的解决方法。首先，选取 StyleGAN 图像生成模型作为研究对象，并分析 StyleGAN 生成模型内部不同的隐编码空间。由于人脸图像中具有丰富的语义信息，例如五官和表情等。为了方便图像编辑任务的研究，我们选取在人脸生成任务上公开的 StyleGAN 预训练模型。StyleGAN 模型作为图像生成模型具有良好的生成效果，内部具有现成

的 Style 编码空间 $\mathcal{W}$。GANSpace 采用 PCA 方法分析高斯噪声空间 $\mathcal{Z}$ 和 Style 空间 $\mathcal{W}$ 。对比原始的高斯噪声空间 $\mathcal{Z}$，Style 编码空间 $\mathcal{W}$ 更加解耦使其容易发现可分离的语义信息。StyleSpace 将 Style 空间映射到合成网络的不同生成层的中，提出了分层分布的 $\mathcal{S}$ 编码空间并解决了单通道的图像属性编辑问题。对比 Style 编码空间 $\mathcal{W}$，分层的 $\mathcal{S}$ 编码空间在语义编辑上能获取更解耦的编辑向量。其次，提出了指定编辑属性内容去寻找特征编辑向量的方法。为获取指定的人脸属性，我们选取了人脸语义分割模型和内容属性分类模型来表示图像中的语义属性。从知识蒸馏学习方法角度出发，我们引入了区域和内容属性的预训练模型作为教师模型。预训练的 StyleGAN 模型成为待分析的学生模型。教师模型用于从 StyleGAN 模型中分析不同编码对特定属性的响应，并按照响应度对编码通道进行排序，选择 Top-K 作为响应通道。具体而言，对于图像中指定的区域，选取了预训练的语义分割神经网络模型，计算的编码对指定属性的梯度。梯度的大小反应了不同编码通道对指定属性的响应，可以通过筛选绝对值最大的梯度来解耦出对应属性的编码通道。然而，在 StyleGAN 模型的不同层的隐特征编码空间中，我们发现不同层间的编码通道进行梯度比较，无法做出合理解释。因此，在分层的 $\mathcal{S}$ 编码空间上，我们提出了先过滤指定生成层再筛选响应通道的方法。最后，我们量化了单通道和多通道的属性编辑强度。对于指定的属性编辑，可以定位到单个的响应通道，并利用正负样本在该通道的均值和方差，以此确定单通道的单位编辑强度和范围。对于多通道编辑方法，我们可以归一化梯度信息确定多通道间的联合编辑强度但无法确定可编辑的范围。本章的主要贡献如下所示：

- 针对 StyleGAN 模型，提出了一种采用逐层梯度分析的新检测方法，用于定位指定属性的编码通道。基于知识蒸馏学习方法，我们将属性属性预训练模型作为教师模型，用于定位指定属性的特征编码通道。通道定位方法利用属性预训练模型计算每个编码通道的梯度，并进行分层排序筛选，选取对属性响应较高的编码通道。
- 利用提出的编码通道检测方法，在分层的 Style 编码空间 $\mathcal{S}$ 上提出了单通道和多通道特征编辑方法，并量化了单通道和多通道操作的编辑强度。
- 在人脸编辑实验中，完成了各种指定属性的可控编辑，充分体现了该方法在细粒度图像编辑上的优越性。对比其他编辑方法，我们提出的基于特征编码的图像编辑方法实现了多种指定属性的内容编辑。结合图像特征编码的研究，我们实现了对真实人脸图像的编辑。

6.2 基于 StyleGAN 模型的可控编辑方法

6.2.1 图像特征编码空间及编辑方法

从图像特征编码的角度分析，StyleGAN 网络实现了从 512 维的高斯噪声编码 $\boldsymbol{z}$ 到 1024×1024 维图像的生成，其中噪声编码空间 $\mathcal{Z}$ 可以看作为显式的图像特征编码。映射网络将噪声编码映射到 512 维的 Style 编码，构成了隐编码空间 $\mathcal{W}$。对比噪声编码空间 $\mathcal{Z}$，隐编码空间 $\mathcal{W}$ 具有更好的解耦性，编码通道对应着不同语义属性。按照合成网络的分层结构划分，Style 编码 $\boldsymbol{w}$ 被映射成 18×512 维的 $\boldsymbol{w}$+ 编码和 26×512 维的 $\boldsymbol{s}$ 编码。$\mathcal{W}$+ 和 $\mathcal{S}$ 编码空间比 $\mathcal{W}$ 空间更加庞大且复杂，但能进一步解耦特征编码，分离图像中不同的语义内容。表 6-1展示了 StyleGAN 网络中不同的隐编码空间。表 6-2详细介绍了 StyleGAN 合成网络中不同的生成层所对应的编码向量 $\boldsymbol{w}$+ 和 $\boldsymbol{s}$。由于隐特征编码看作为一种图像特征，关于语义内容具有良好的的可解释性，本章主要研究基于隐特征编码的图像编辑方法。

聚焦于预训练 StyleGAN 模型的隐特征编码，不同隐空间下的编辑方法会影响图像的编辑效果。根据第5章的研究，StyleGAN 中的 Style 编码被作为一种图像特征，被成功应用于图像编辑等任务中，例如 GANSpace[40] 方法。然而，对于人脸中指定属性的内容编辑，Style 编码所在的 $\mathcal{W}$ 空间的编辑向量通常是耦合的。Style 编码与编辑的语义内容难以准确地对齐。当修改 Style 编码时，不仅指定的语义属性会发生改变，图像中其余内容属性也会发生迁移。这种情况在 $\mathcal{W}$ 空间中普遍存在，这十分影响图像编辑的质量。其中，主要原因在于 Style 编码无法做到控制单一的特征表达。由于 $\boldsymbol{w}$ 编码实际在合成网络中常以 $\boldsymbol{w}$+ 的形式应用于合成网络的生成中，StyleSpace[62] 分析了在 $\mathcal{W}$ 和 $\mathcal{S}$ 空间上的分层特征编码，并验证了分层的特征编码比 Style 编码更加解耦，能有效表达单一的图像语义内容。在 $\mathcal{S}$ 空间上，StyleSpace 提出了对于特定人脸的单通道编码的定位和编辑。受上述工作的启发，本章主要围绕 $\mathcal{S}$ 编码空间上进行内容及语义属性的定位和编辑。

基于预训练的 StyleGAN 模型在人脸生成上的研究，本章主要面向人脸属性编辑问题，并提出特征编辑方法实现指定人脸属性的修改。图 6-1可视化了所提出的方法，其中主要包含有属性模型（教师模型），合成网络（学生模型）和属性编码的定位及编辑方法。首先，利用教师模型，明确了人脸属性的

表 6-1: StyleGAN 网络模型中不同隐特征空间的编码维度。

特征编码空间	$\mathcal{Z}$	$\mathcal{W}$	$\mathcal{W}$+	$\mathcal{S}$
编码维度	512	512	18×512	26×512

可编辑内容。人脸的编辑内容主要包括有五官区域和表情等。基于知识蒸馏学习方法，我们引入了额外的属性教师模型，用于指定编辑的内容。属性模型来自于其他图像任务的预训练模型，如人脸语义分割模型和属性分类模型。当属性模型为预训练的人脸语义分割模型时，可以指定属性编辑区域包括有头发，眼睛和嘴巴等。然而，这种指定区域的属性模型难以区分更细粒度的语义特征，如头发区域内的发型和发色的差异。因此，还引入了人脸的语义分类模型，用于更细粒度地指定编辑内容，如红黑发色，波浪发型，山羊胡子等语义属性。其次，在特征编码空间中发掘指定属性的编辑方向或通道。为了查明编码向量空间 $\mathcal{S}$ 中潜在的编辑方向，需要在特征编码通道间对指定的属性内容进行定位。固定预训练的合成网络和属性分类模型的参数，将特征编码 $\boldsymbol{s}$ 看作为待研究的变量。设特征编码 $\boldsymbol{s}$ 对应生成图像 I。利用指定的属性分类网络，计算生成图像 I 对指定属性的分类损失，并反向传播计算该图像对该属性的梯度。生成图像的梯度表示该像素对于指定属性的响应。梯度的绝对值越大代表该像素与指定属性越相关。利用生成图像的梯度，在合成网络中链式传播图像的梯度信息，计算特征编码 $\boldsymbol{s}$ 对于指定属性的梯度。由于特征编码 $\boldsymbol{s}$ 在生成网络中分层表示，我们首先选取与指定属性相关的编码层，再排序筛选响应的特征编码通道。最后，选取指定属性的编码通道后，还量化了特征编码编辑扰动的方向和可编辑距离，实现可控的指定内容的修改。我们分别提出了单通道和多通道的特征编辑方法，以应对众多的人脸属性。从模型间的学习方法角度分析，本文所提出的指定属性的特征编辑方法属于知识蒸馏学习方法。通过教师模型实现对属性编码通道的定位，从而学习特征编码地编辑向量和强度范围。

6.2.2 基于生成区域的编码通道定位及编辑

首先，研究了基于生成区域的特征编码的定位和编辑方法。假设生成图像 I 由特征编码 $\boldsymbol{s}$ 通过预训练的 StyleGAN 的合成网络所生成，即 $I = G(\boldsymbol{s})$。特征编码 $\boldsymbol{s}$ 来自于 Style 编码在合成网络 G 中进行分层变换的结果。我们试图在分

表 6-2: 按分层结构划分的编码空间 $\mathcal{W}+$ 和 $\mathcal{S}$。

$\mathcal{W}+$	$\mathcal{S}$	# Channels with $\mathcal{S}$	Output resolution
$\mathbf{w}^0$	$\mathbf{s}^0$	512	4×4
$\mathbf{w}^1$	$\mathbf{s}^1$	512	4×4
$\mathbf{w}^1$	$\mathbf{s}^2$	512	8×8
$\mathbf{w}^2$	$\mathbf{s}^3$	512	8×8
$\mathbf{w}^3$	$\mathbf{s}^4$	512	8×8
$\mathbf{w}^3$	$\mathbf{s}^5$	512	16×16
$\mathbf{w}^4$	$\mathbf{s}^6$	512	16×16
$\mathbf{w}^5$	$\mathbf{s}^7$	512	16×16
$\mathbf{w}^5$	$\mathbf{s}^8$	512	32×32
$\mathbf{w}^6$	$\mathbf{s}^9$	512	32×32
$\mathbf{w}^7$	$\mathbf{s}^{10}$	512	32×32
$\mathbf{w}^7$	$\mathbf{s}^{11}$	512	64×64
$\mathbf{w}^8$	$\mathbf{s}^{12}$	512	64×64
$\mathbf{w}^9$	$\mathbf{s}^{13}$	512	64×64
$\mathbf{w}^9$	$\mathbf{s}^{14}$	512	128×128
$\mathbf{w}^{10}$	$\mathbf{s}^{15}$	256	128×128
$\mathbf{w}^{11}$	$\mathbf{s}^{16}$	256	128×128
$\mathbf{w}^{11}$	$\mathbf{s}^{17}$	256	256×256
$\mathbf{w}^{12}$	$\mathbf{s}^{18}$	128	256×256
$\mathbf{w}^{13}$	$\mathbf{s}^{19}$	128	256×256
$\mathbf{w}^{13}$	$\mathbf{s}^{20}$	128	512×512
$\mathbf{w}^{14}$	$\mathbf{s}^{21}$	64	512×512
$\mathbf{w}^{15}$	$\mathbf{s}^{22}$	64	512×512
$\mathbf{w}^{15}$	$\mathbf{s}^{23}$	64	1024×1024
$\mathbf{w}^{16}$	$\mathbf{s}^{24}$	32	1024×1024
$\mathbf{w}^{17}$	$\mathbf{s}^{25}$	32	1024×1024

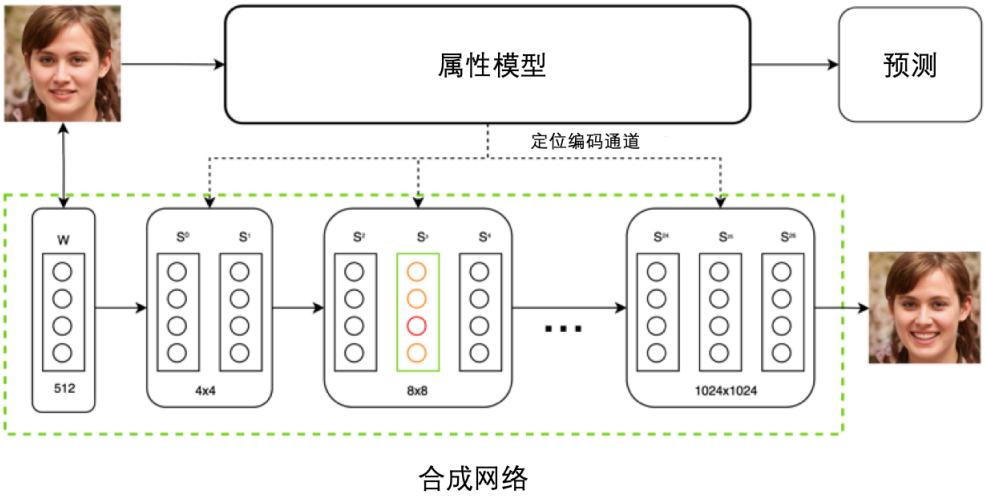


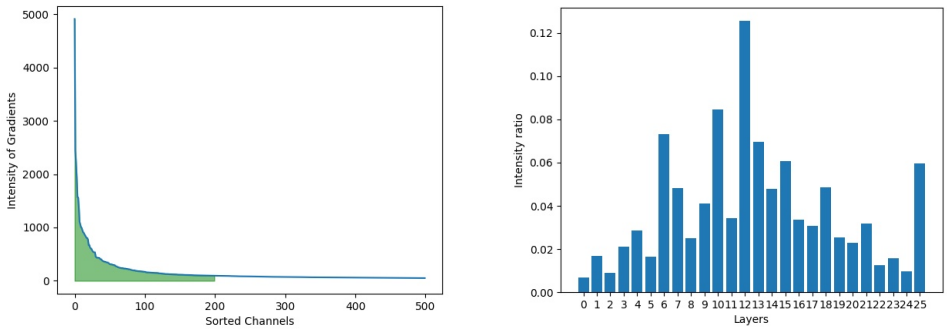
图 6-1: 指定属性内容的人脸编辑方法的流程图。其中，属性模型（Attribute Model）用于指定待编辑的属性。图像特征通道的定位方法用于在合成网络（Synthnsis Network）的分层特征编码中选择负责指定属性生成的通道。

层结构的合成网络中对 $\boldsymbol{s}$ 编码进行指定属性内容的定位，并对其进行编辑实现图像内容的细粒度修改。在人脸区域的编辑问题中，采用预训练的语义分割网络模型来预测人脸图像 I 中五官的语义掩码 M。对于指定的局部目标区域 r，例如嘴巴，鼻子等，获取该语义掩码区域为 M^r。为了获取特征编码对该区域的响应，计算该指定区域 M^r 上的每个像素对特征编码 $\boldsymbol{s}$ 的平均梯度，表示为：

$$\mathbf{g}_{\boldsymbol{s}}^{r} = \frac{1}{\sum_{p \in \mathbf{M}^r}} \sum_{p \in \mathbf{M}^r} \frac{\partial \mathbf{I}_p}{\partial \boldsymbol{s}}, \tag{6-2}$$

其中 p 为每个像素的二维坐标。$\mathbf{g}_{s_j^i}^r$ 表示对第 i 个生成层中第 j 个通道对于人脸区域 r 的平均响应。$\mathbf{g}_{s_j^i}^r$ 的强度越高代表该通道与该区域的关联度越高。该通道可能与该区域内的语义内容生成有关。图 6-2(a) 显示了 $\mathbf{g}_{\boldsymbol{s}}^r$ 在“嘴”区域的前 500 强的通道分布。特征编码通道的梯度 $\mathbf{g}_{\boldsymbol{s}}^r$ 的分布不同于长尾分布，前 200 通道的梯度响应总和占总通道响应总和的 97.3%。这表明了不同生成区域在特征编码空间 S 中是相互可分离的。

然而，由于所研究的 $\mathcal{S}$ 编码空间分布在合成网络的所有生成层中，直接比较不同层间的编码梯度强度并不合理。为了获取对特定属性区域的响应，我们分析了不同生成层的编码在区域内的平均梯度。图 6-2(b)展示了 26 层的 $\boldsymbol{s}$ 编码的平均梯度。进一步分析表明在 StyleGAN 的合成网络中，不同生成层中的编



(a) 关于人脸嘴部区域的的编码通道的梯度响应图。只选取绝对值前 500 大的编码通道。 (b) 关于嘴部特征的分层 $\boldsymbol{s}$ 编码的平均梯度响应。

图 6-2: 分析隐特征编码空间 $\mathcal{S}$ 上不同通道和生成层对指定属性的梯度响应。

码负责不同目标区域及语义信息。对于特征编码 $\boldsymbol{s}$，合成网络中的低分辨率生成层控制生成图像的位置和姿态，中间的生成层控制不同语义内容的生成，最后的高分辨率生成层用于控制全局图像的色彩和阴影等。为了获取负责指定区域响应的编码通道，我们按照 $\boldsymbol{s}$ 编码的层次结构，先筛选出响应最大的层，并在这些层中选取前 K 大的梯度响应的通道作为控制指定区域的关联通道结果。由于 $\boldsymbol{s}$ 中的部分的单编码通道，能有效控制人脸不同语义属性的生成。我们首先提出了单编码通道编辑方法，来确定对指定区域内容的影响，即需要手动分析属性区域中每个通道的语义信息。在操作单通道编辑之前，我们计算平均值和标准差 $\boldsymbol{\delta}$ 的 1000 个样本来设置每个样式通道的单位编辑强度，并将标准差作为单位编辑强度。对于指定区域属性，特征编码 $\boldsymbol{s}$ 的编辑操作被定义为:

$$\boldsymbol{s}^* = \boldsymbol{s} + \alpha \boldsymbol{\epsilon}^*. \tag{6-3}$$

其中，$\boldsymbol{\epsilon}^*$ 为特征编码的单位编辑方向，超参数 α 为单位编辑方向上的距离。

6.2.3 基于语义属性的编码通道定位及编辑

基于生成区域的特征编码定位及编辑方法存在的主要问题是无法进一步编辑区域内的属性内容。该方法只能初步定位不同编码通道对生成内容的影响，但是无法细粒度地定位更加抽象的语义属性。例如，上述方法无法实现对微笑，年龄等语义内容的编辑。在人脸属性的编辑问题上，我们试图完成更加复杂的语义内容编辑。为了实现这一目标，依据图 6-1所介绍的方法流程，我们提出了基于语义属性的编码通道定位及编辑方法。从模型间的学习方法上分

析，该方法与基于生成区域的编辑方法最大的不同在于教师模型。基于语义属性的编码通道定位及编辑方法采用了属性分类教师模型用以细粒度地描述人脸语义属性，但特征通道的定位和编辑方法大致相同。

当以人脸属性分类器作为教师模型时，特征编码可以定位到更加具体的语义属性而不仅仅局限在某个生成区域内。设预训练的语义属性 a 的分类器为网络模型 F_a，对于生成图像 $\mathbf{I}$ 的属性损失定义为交叉熵损失 ℓ_a。依据反向传播的链式法则，$\boldsymbol{s}$ 编码对该属性 a 损失的梯度为：

$$\mathbf{g}_{\boldsymbol{s}}^{a} = \frac{\partial \ell_a}{\partial \boldsymbol{s}} = \frac{\partial \ell_a}{\partial F_a(\mathbf{I})} \frac{\partial F_a(\mathbf{I})}{\partial \mathbf{I}} \frac{\partial \mathbf{I}}{\partial \boldsymbol{s}}. \tag{6-4}$$

由于单张图像的对指定属性的梯度响应存在误差，且当生成图像的内容没有包含属性的内容时无法用于定位响应通道。因此，我们需要利用属性分类器模型筛选出含有该属性的生成图像，并将其作为正样本，对未包含该属性的负样本进行摒弃。为了提高定位及编辑编码方法对指定内容的泛化性能，我们在正样本 $\boldsymbol{s}$ 编码集合中进行编码通道的定位。将 $\boldsymbol{s}_i$ 表示为第 i 个正样本的特征编码，则特定属性的正样本集合表示为 $P = [\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_n]$。我们计算在该正样本集合 P 上编码的累计的平均梯度作为不同通道对该属性的响应为：

$$Average(\mathbf{g}_{\boldsymbol{s}_i}^{a})_{\boldsymbol{s}_i \in P} = \frac{1}{|P|} \sum_{\boldsymbol{s}_i \in P} \mathbf{g}_{\boldsymbol{s}_i}^{a}. \tag{6-5}$$

同时，依据层筛选机制，获取指定属性 a 的响应生成层 l。再选取前 K 大的 $Average(\mathbf{g}_{\mathbf{s}_i}^{a})_{\mathbf{s}_i \in P}$ 作为编码通道对指定属性的响应。则对于属性 a 所筛选出的第 l 层特征编码的前 k 大的响应通道集合为 C_l^k，表示为：

$$C_l^k = \{\text{Top-}k \text{ channels on } Average(\mathbf{g}_{\mathbf{s_i}}^{a})_{\boldsymbol{s}_i \in P}\}. \tag{6-6}$$

其中 $Average(\mathbf{g}_{\boldsymbol{s}_i^l}^{a})_{\boldsymbol{s}_i \in P}$ 表示为计算正样本集合编码的第 l 层对属性 a 的平均响应。

当定位到指定语义属性的响应通道后，我们也提出了单通道和多通道的编辑方法。单通道的编辑方法与式6-3类似，在正样本上计算该通道的均值和方差，并将标准差作为单位编辑强度。然而，对于复杂的语义内容修改，单通道编辑由于只修改其中单个编码通道，难以完整控制整个语义内容。例如，人脸眼睛内容可以由分布在不同层的特征编码 $\boldsymbol{s}$ 联合表示。因此，我们提出了多通道编辑方法。不同于单通道编辑方法，多通道编辑方法的难点在于确定多个通道间的编辑方向。我们首先研究的是同一生成层内的编码的多通道编辑技术。

对于正样本集合统计的平均梯度 $Average(\mathbf{g}^a_{\boldsymbol{s}_i})_{\boldsymbol{s}_i \in P}$ 中，在第 l 层生成编码中，我们归一化前 k 层的平均梯度作为多通道编辑的单位编辑方向，表示为：

$$\boldsymbol{\epsilon_m} = \text{Normalize}(\mathbf{g}^a_{\mathbf{s}_i, C^k_l})_{\mathbf{s_i} \in P}. \tag{6-7}$$

结合式6-3的编辑方法，利用单位编辑方向 $\boldsymbol{\epsilon_m}$ 完成指定属性的多通道编辑。然而，该方法无法确定在该编辑方向上的作用距离。当编辑强度过大时，会造成生成图像失真。此外，由于不同生成层的通道响应无法直接比较，上述方法只适用于单特征通道的筛选，多生成层的多通道编码技术有待进一步的研究。

6.2.4 图像编辑的学习方法

对比基于生成区域的图像特征编辑方法和基于语义属性的图像特征编辑方法，这两种方法都部分实现了指定属性的图像修改，其主要的区别在于可选择属性。生成区域的属性通过语义分割模型指定，语义属性则通过属性分类器指定。从知识蒸馏学习方法角度分析，我们将指定编辑属性的模型看作为教师模型。从学习方法的角度，上述提出的两种特征编辑方法本质相同。然而，上述方法中却没有待学习的学生模型。在 $\mathcal{S}$ 空间的特征编码通道定位方法没有修改 StyleGAN 合成网络的参数，特征通道编辑也作用于分层的 Style 编码 $\boldsymbol{s}$ 上。但可以将 StyleGAN 看作为一种特殊的学生模型，该学生模型试图研究内部特征的图像生成过程。虽然上述方法中没有明确的学生模型，但可以利用教师模型的指导学习 StyleGAN 学生模型中的特征编码与指定属性间的映射关系。即学生模型不仅仅局限于待训练的网络模型，也可以指代从预训练的模型中学习相关知识。此外，本文试图实现对真实图像的编辑。结合第4章对图像特征编码的研究，我们将 StyleGAN 作为教师模型，用于指导训练编码学生模型。即编码学生模型用于提取图像的分层 Style 特征编码，并用于后续的图像编辑。

6.3 实验与分析

本章实验试图控制特征编码实现细粒度的人脸属性编辑。对于任意一张人脸图像，本章综合了运用了上述的研究内容，用以实现可指定属性的人脸编辑。首先，利用点云对齐方法，将人脸对齐到标准的人脸姿态。其次，运用预

训练的特征编码模型，将人脸映射到特征编码。本章采用 StyleGAN 网络的预训练编码提取人脸编码。我们提出了可指定属性的人脸特征编码定位和修改方法，可实现对特征编码连续的细粒度编辑。最后，利用人脸特征解码研究，通过对特征编辑实现人脸图像生成内容的控制。

6.3.1 数据集及预训练模型

我们选择在 StyleGAN 的人脸生成模型上验证提出的编辑方法。StyleGAN 在 FFHQ（Flickr-Faces-HQ Dataset）[92] 数据集上进行对抗学习，得到预训练的人脸生成模型。基于此预训练模型，在高斯噪声中随机采样生成 1000 张图像，并分别获取不同的隐编码向量集 Z, W, S，用于定位属性的编码通道。对于属性教师模型，则选取 BiSeNet[98] 作为预训练的语义分割模型，用于分割不同部位，生成如眼睛，鼻子，头发等区域的语义掩码。同时，也在属性分类数据集 CelebA[96] 上，训练以 ResNet50[1] 为骨架的图像属性分类器。我们总共训练了 40 种描述人脸语义属性的二分类器。

6.3.2 评价指标

对于基于生成区域的编辑方法，主要通过可视化的方法进行定性的研究。对于基于语义属性的编辑方法，我们不仅通过可视化进行定性研究，还对特征编码编辑的效果进行了定量分析。借鉴 StyleSpace 中的的评价方法，可以选取属性分离度（Attribute Dependency，AD）作为评价特征编码对指定属性的解耦程度 [62]。属性分离度评估了进行编码通道编辑后生成图像与原图像在指定属性上的语义距离。当存在 K 种预训练的属性分类器集合 $\mathcal{A} = [a_0, ..., a_k]$ 时，设第 i 个属性分类器对于输入图像的 logit 输出为 l_i，则编辑后的图像与原图像的 L_2 距离为 Δl_i。对指定属性 a_t 进行编辑，在单位编辑方向上属性分离度为 $AD_t = \Delta l_t$，用于衡量指定属性 a_t 的编辑图像与原图像的修改距离。对于不期望修改的属性，属性分类度为 AD_o，表示在其余属性上的编辑距离为 $\frac{1}{k}\sum_{i\in\mathcal{A}\setminus a_t}\left(\frac{\Delta l_i}{\sigma(l_i)}\right)$，其中 $k = |\mathcal{A}| - 1$。当进行指定属性修改时，我们希望所编辑的编码通道只对选取的属性 a_t 产生影响，而不对其余的属性做相关改动。同时，AD 分离度距离容易受到特征编辑强度的影响。不同的方法对编码通道的单位编辑强度存在不一致，并且不同属性的编辑强度也存在差异。我们计算了更稳定的指标 $\frac{AD_t}{AD_o}$，用于表示期望编辑属性和不期望编辑属性的比值。这些评价指

图 6-3: 基于生成区域的图像编辑的可视化结果。

标在生成的 1000 张测试样本集上进行评估。通过属性分类器方法选取 20-50 个正样本，用于特征编码的通道的定位，并设置单位编辑向量和编辑强度范围。

6.3.3 图像编码编辑实验和分析

基于内容区域的编辑结果：图 6-3展示了对不同人脸区域进行编码编辑的生成结果，主要包括有头发，嘴部，眼睛等区域。在图 6-3中，第一列图像来自于预训练的 BiSeNe 教师模型对同一张图像不同部位的语义分割图。第二至四列为利用式 (6-2) 中提出的定位方法，通过定位编码通道实现与指定语义区域进行关联，并通过式 (6-3) 进行单通道编辑的结果，其中左上角的红色数字来表在合成网络中定位到的层和通道。例如，“6-188”表示 $\mathcal{S}$ 编码空间中的第 6 生成层的第 188 位通道。从图 6-3中可以看出，对于指定人脸内容区域，通过第 6.2.2小节提出定位编码通道及编辑方法，我们实现了对该区域的内容编辑。$\mathcal{S}$ 编码通过分布在不同层的编码通道，控制区域内容生成。合成网络分层次地利用 $\boldsymbol{s}$ 编码，由粗粒度到细粒度的方式进行渲染，生成伪造图像。

基于语义属性的编辑结果：图 6-4展示了对人脸中不同语义属性进行单通道编辑的生成图像，语义信息来自于 CelebA 所标注的 40 种属性类别。从

表 6-3: 关联指定语义属性的响应层和通道。

Attribute	Old / Young	Goatee	Hairline	Smiling
(layer, channel, rank)	(9, 435, 1)	(9, 421, 1) (9, 6, 2) (12,237,3)	(6, 322, 1) (6, 504, 2) (6, 364, 3)	(6, 501, 1) (6, 113, 2) (6, 378, 4)
Attribute	Eyeglass	Arched	Big Nose	Female / Male
(layer, channel, rank)	(3, 228, 1) (3, 120, 2) (2, 175, 3)	(6, 35, 1) (9, 340, 2)	(6, 501, 1) (6, 110, 2)	(9, 6, 1)
Attribute	Lipstick	Wavy Hair	Narrow Eye	Colour Hair
(layer, channel, rank)	(15, 45, 1)	(9, 475, 1) (6, 323, 1) (6, 500, 3)	(11, 257, 1) (9, 63, 2) (14, 239, 3)	(11, 286, 1) (12, 424, 1) (15, 62, 1)

图 6-4可以看出对于特定的属性，所提取的编码通道在图像编辑上具有一致性。即单通道编辑方向可以在不同图像中实现一致的属性篡改。这使得通过特征编码快速实现任意图像的指定内容修改成为可能。在人脸属性编辑问题上，我们提出的方法在可编辑的语义属性的数量上显著优于其他方法。例如，InterFaceGAN 和 StyleFlow[41] 方法发现了 5 种人脸属性的特征编码的编辑方法。Siavash 等人 [99] 利用标注的属性样本，学习属性嵌入特征编码网络，实现了 35 种人脸属性编辑。但我们提出的方法的可编辑的属性数量受限于预训练的属性分类器本身，具有良好的泛化性。同时，该方法能够快速且有效地定位出多种与指定属性关联的特征编码，适用于多种语义属性的编辑。表 6-3总结了不同语义属性在不同生成层所定位到的编码通道。有些属性可以只有单通道控制，例如年龄，性别等，但有些属性具有多种表达，不同通道对应更细粒度的区分，如发型，发色等。属性通道的定位研究部分解释了部分编码的细粒度语义内容。

与其他方法的对比结果：在隐编码空间 $\mathcal{S}$ 上，我们尝试定量地比较了在不同属性上的单通道和多通道编辑方法的结果。同时，我们计算提出的编辑方法在生成图像上的属性分离度，并定量对比在 StyleSpace 中的最先进的编辑方法。对于指定属性的修改，表 6-4比较了各种编辑方法在生成图像上的属性分离度。表 6-4中的实验数据来自上述随机采样的 1000 张图像中的正样本的检测

图 6-4: 基于语义属性的图像编辑的可视化结果。

结果。在表 6-4中，目标属性分离度 AD_t 越高越好，其余属性分离度 AD_o 越低越好。当 $\frac{AD_t}{AD_o}$ 的比值小于 1 时，则代表该方法未能检测到分离的编码通道，无法控制指定属性内容的修改。例如，StyleSpace 方法无法实现在 Eyeglasses 和 Age 属性上的编辑。在隐编码空间 $\mathcal{S}$ 上，我们提出的方法比 StyleSpace 方法能更加准确地定位出的不同属性对应的特征编码通道。在可编辑属性的质量和数量上，提出的单通道和多通道编辑方法也优于 StyleSpace 方法。

表 6-4: 对比不同特征编辑方法在各种语义内容上的属性分离度。

Methods	Eyeglasses	Goatee	Smiling
	$(AD_t, AD_o, \text{Ratio})$	$(AD_t, AD_o, \text{Ratio})$	$(AD_t, AD_o, \text{Ratio})$
StyleSpace	(0.40, 0.76, 0.53)	(2.76, 0.50, 5.49)	(2.94, 1.19, 2.46)
Single-channel	(2.61, 0.35, 7.31)	(2,36, 0.38, **6.20**)	(5.67, 0.67, 8.42)
Multi-channel	(5.28, 0.71, **7.4**)	(6.72, 1.19, 3.50)	(7.80, 0.88, **8.88**)
Methods	Gender	Black Hair	Age
	$(AD_t, AD_o, \text{Ratio})$	$(AD_t, AD_o, \text{Ratio})$	$(AD_t, AD_o, \text{Ratio})$
StyleSpace	(2.54, 1.34. 1.89)	(4.38, 0.54, 8.03)	(5.29, 6.76, 0.78)
Single-channel	(5.08, 1.12, **4.54**)	(4.38, 0.54, 8.03)	(4.32, 1.65, **2.54**)
Multi-channel	(5.95, 1.36, 4.39)	(10.05, 1.04, **9.61**)	(3.92, 2.74,1.43)

6.3.4 消融实验和超参数分析

梯度分层筛选的消融实验：为验证提出方法对于编码进行分层筛选的必要性，我们进行了相关的消融实验并进行分析。设置的对比方法包括有直接基于梯度的编码编辑方法，基于前 k 大的梯度筛选的通道编辑方法和所提出的分层提出前 k 大的单通道编辑方法。不同的特征编辑方法主要区别在于对指定属性的响应通道的定位上。图6-5可视化展示了利用不同梯度筛选方法，定位到的通道并进行编辑的结果。其中，第一列为原始的生成图像。第二列为利用分层梯度筛选定位的单通道的编辑结果，第三列为未分层的编码前 $\boldsymbol{K}$ 大的梯度的多通道编辑结果，第四列为直接归一化编码 $\boldsymbol{s}$ 作为单位编辑向量所生成的图像。对比第二列和第四列的结果，利用指定属性的编码梯度来修改图像特征编码 $\boldsymbol{s}$，虽然可以编辑指定属性的内容，但也显著修改了其余不相关的属性，导致编辑后的图像严重失真。对比第二列和第三列的结果，当在所有层中进行前 k 大的梯度筛选通道时，由于不同层间的梯度无法直接进行比较，导致丢失了对属性编辑影响最大的通道检测，无法实现指定内容的编辑。因此，我们提出的分层提取筛选并提取前 k 大梯度的通道检测方法，能够在 $\boldsymbol{s}$ 编码中实现快速且精确的指定属性的定位，并改善语义内容的编辑结果。

可控的连续编辑分析: 由于在单通道编辑方法中量化了单位编辑方向，结合式 (6-3), 我们通过控制超参数 α 实现特征编码 $\boldsymbol{s}$ 在指定语义属性的连续

图 6-5: 在人脸的 “胡子” 属性上，基于不同梯度筛选方法所检测的编码通道的编辑结果。

编辑。图 6-6展示 InterfaceGAN 和提出的单通道编辑方法在人脸年龄属性上的连续编辑结果。其中，InterfaceGAN 实现了在隐特征空间 $\mathcal{W}$ 上的连续编辑，单通道编辑方法通过修改超参数 α 实现在特征编码空间 $\mathcal{S}$ 上的连续编辑。如图 6-6所示，当沿着年龄属性的编辑方向增加编辑强度时，InterfaceGAN 将逐渐暴露出所检测的编辑方向在编码空间中解耦不充分的缺点，其所获取的年龄属性编辑方向与眼睛属性内容耦合。对比 InterfaceGAN 方法，我们提出的单通道检测方法在可控的连续编辑上保持一致性。在图 6-6中，当增加沿年龄属性的编辑强度时，生成人脸的脸部皱纹的逐渐增加，体现了在年龄上的连续性。同时，再连续修改的年龄的特征编码后未引起其余特征的显著修改，显示了在 $\mathcal{S}$ 编码空间上特征与语义内容的高度解耦。

真实图像上的特征编码编辑方法的研究：当给定真实的人脸图像时，我们希望利用提出的编辑方法，对其进行指定语义内容的编辑。由于本章所提出的方法是基于 StyleGAN 中的隐特征编码空间 $\mathcal{S}$，无法直接应用于真实的图像编辑中。根据第 4章和第 5章到的研究成果，我们采用了一种简单的解决方案。首先，学习图像到 Style 编码的编码器，将图像映射到 Style 特征编码。这部分的工作可以采用 Edit4Edit 方法 [63] 所学习的编码网络。其次，利用合成网络进行解码。将从图像中提取到的 Style 编码映射到特征编码 $\boldsymbol{s}$。最后，基于特征编码的编辑方法，指定待修改的属性。当完成各种属性的编码通道的定位后，可以直接对该特征编码 $\boldsymbol{s}$ 进行编辑。上述流程实现了在真实人脸图像上指定属性的编辑。如图 6-7所示，可视化展示了在真实人脸图像上的各种属性编辑的结果。从侧面验证了提出的通过修改特征编码编辑指定图像内容的可行性。

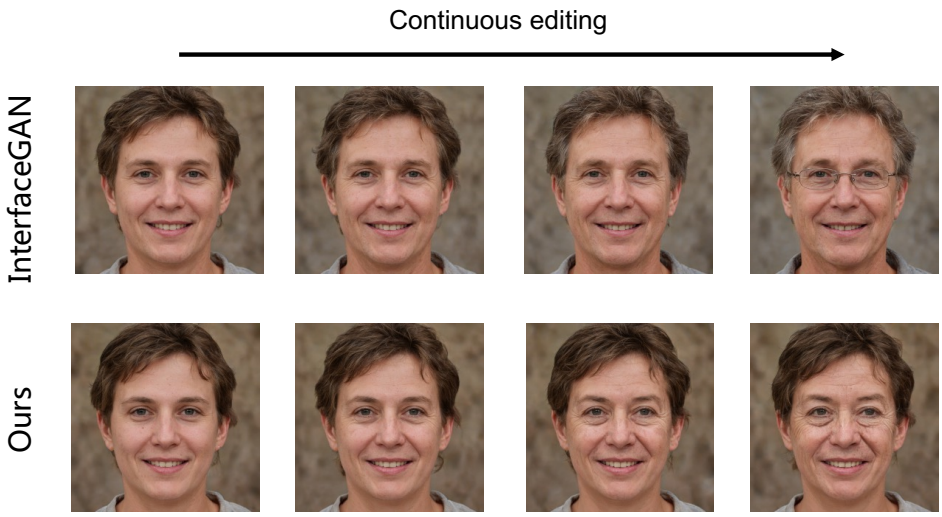


图 6-6: 对比 InterfaceGAN 和提出的单通道编辑方法在“年龄”属性上的的连续编辑结果。从左至右表示为维持上述方法的编辑方向不变，编辑强度逐渐增强的生成图像。

6.4 本章小结

本章主要研究了图像生成模型中的隐特征编码的编辑问题，通过编辑特征编码实现指定语义内容的可控生成。本章主要基于预训练的人脸生成 StyleGAN 模型，提出了可指定修改属性的编码通道定位及编辑方法。在人脸编辑问题中，为指定修改的属性内容，将预训练的人脸语义分割模型和人脸属性分类模型作为教师模型，用于定位 StyleGAN 生成模型中与其关联的隐特征编码通道。我们提出了将特征编码对教师模型的梯度作为对指定属性的响应。编码对指定属性的梯度越大，代表与其存在密切的关联，可能控制该属性的生成。为了解决不同生成层中编码的梯度难以比较的问题，首先通过分层的梯度筛选负责属性生成的层。之后在此响应层上选取与指定属性编辑相关的编码通道。我们的方法能够快速利用少量的正样本实现对属性的编码通道的定位。最后，还提出了单通道和多通道的编辑方法，并实现了基于内容区域和语义分类属性的人脸编辑。

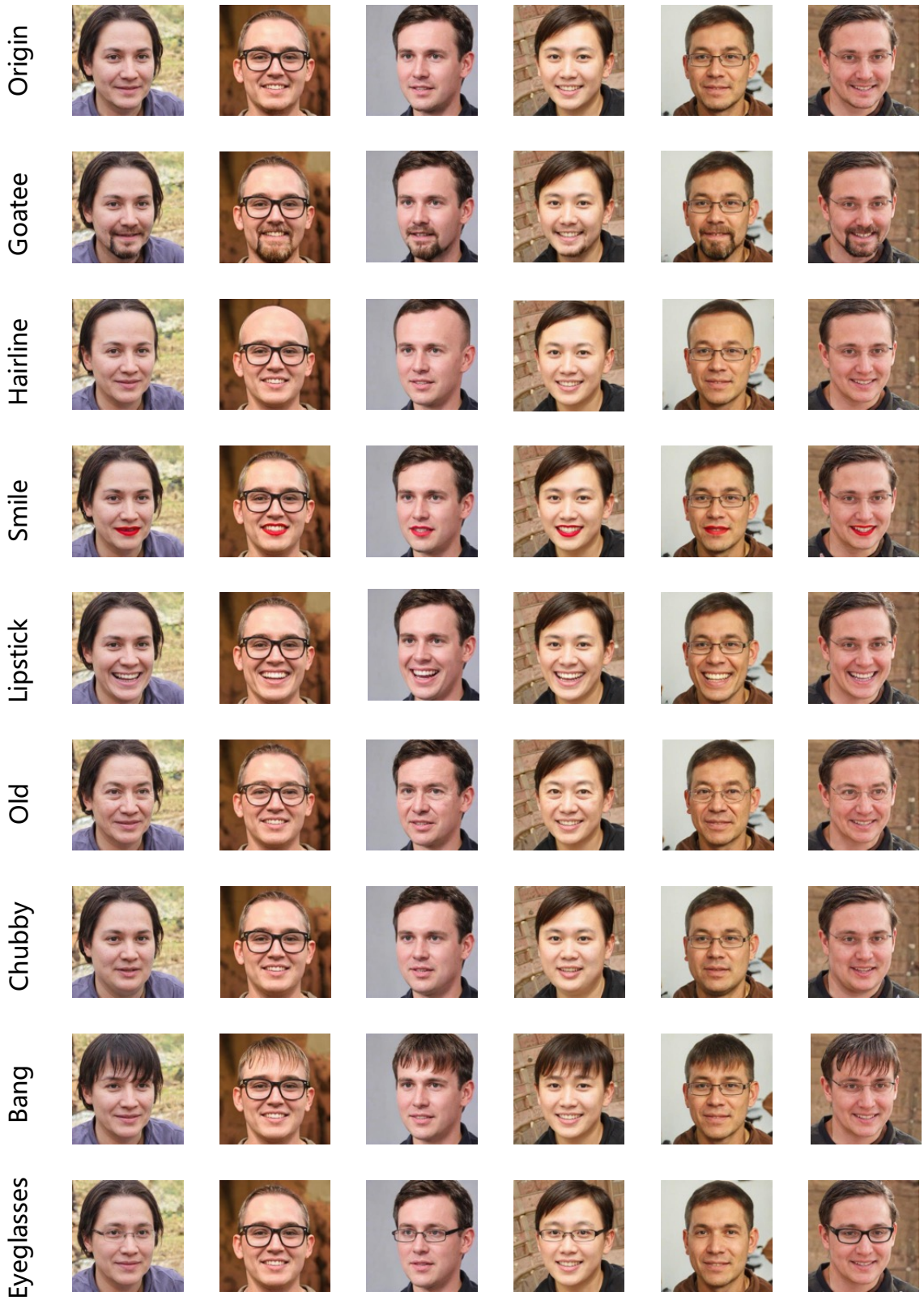


图 6-7: 对真实人脸图像进行多种属性内容编辑的可视化结果。人脸图像的 Style 特征编码提取于 Edit4Edit 方法 [63]。

第七章　总结与展望

7.1　全文总结

本文围绕人脸图像编辑任务系统性地开展了多方面的研究。针对图像编辑问题中核心的研究内容，即图像特征，我们分别研究了图像特征的编码、解码和编辑等内容。首先，为了实现端到端的人脸图像编辑，先研究了局部点云对齐问题，并用于处理人脸图像的对齐。这些工作可被用于人脸编辑的预处理过程，提升图像编辑的质量。其次，本文将人脸编辑中涉及图像特征的工作划分为特征编码和解码两个阶段，并分别独立地研究图像特征编码和解码过程。最后，为了进一步理解图像特征编码的语义内容，我们还研究了针对图像特征的可控编辑，实现了人脸特征编码和解码的统一。本文提出并解决了可指定属性的人脸编辑任务，丰富了人脸编辑的研究成果。本文的具体研究内容和贡献有以下四点：

1. 图像对齐旨在统一图像的编解码过程，是后续进行人脸图像编辑的预处理步骤。我们解决了局部点云的对齐问题，并将其应用都人脸图像编辑问题的对齐预处理中。由于图像编解码被相对独立的研究，利用局部点云的对齐任务，端到端地实现输入图像到生成图像中语义内容的对齐。人脸编辑主要针对标准姿态的人脸进行处理。对于任意获取的人脸图像，将其对齐到标准姿态的预处理步骤，有利于上述人脸图像的编解码过程。我们研究了成对局部点云的对齐及补全任务，并提出了基于混合优化的点云对齐方法。该方法利用无约束变量在整个变换矩阵空间进行迭代优化，解决了点云的局部和全局匹配，提高了对齐精度。该研究可以辅助进行人脸图像对齐，为后续人脸编辑进行铺垫。
2. 针对人脸特征编码问题，本文研究了从无标记数据中学习图像特征编码。为了评估编码网络学习到的图像特征编码的优劣，我们将特征编码应用到图像聚类任务中，以深入研究无标注图像数据下的特征编码的学习问题。为了解决图像聚类任务，我们提出了 DERC 聚类方法，实现了无标注图像上的特征编码的学习。其中，从数据角度分析，DERC 方法采用自编码器模

型完成图像到特征、特征到图像的学习。从模型角度分析，DERC 方法通过编码器和解码器进行合作学习，提取图像中的特征编码。DERC 方法将传统聚类方法作为教师模型，实现对编码网络的指导学习。该研究初步研究了人脸编辑问题中图像的特征编码。

3. 图像特征解码是理解抽象图像特征编码的有效方法，也是解决各种图像任务的关键步骤。本文面向基于特征的图像生成解码研究，并主要围绕预训练的 StyleGAN 生成模型展开。首先，阐明 Style 编码属于一种特殊的图像特征编码。对比随机高斯噪声编码，Style 编码有效提高了图像生成的质量。其次，基于模型间的学习方法，梳理了 Style 编码的学习方法。将 StyleGAN 模型分解为生成模型 G 和判别模型 D，并采用对抗损失进行两个学生模型的生成对抗学习。同时，将生成模型 G 可以进一步分解为映射网络 M 和合成网络 S。StyleGAN 把预训练的 VGG 图像特征编码网络作为教师模型，用于评估生成图像的语义内容差异，并指导映射网络 M 的学习。最后，实验验证了 Style 编码在图像特征解码过程中的优越性。该研究初步分析了基于 Style 编码的合成网络在人脸生成时的解码过程，启发了基于 Style 编码的图像编辑解码方法的研究。

4. 图像特征编码和语义内容存在对应关系。通过对图像特征编码和解码的研究，我们进一步探索了基于图像编码的可控编辑，实现对图像内容的修改。在人脸编辑任务中，我们分析了预训练的 StyleGAN 模型的编码空间，提出了可指定修改内容的编码通道定位及编辑方法。为了指定待编辑的语义属性，将预训练的人脸语义分割模型和人脸属性分类模型作为教师模型，并用于定位合成网络中分层次的特征编码通道。特征编码通道的定位方法是学生和教师模型的知识蒸馏学习方法。我们将特征编码对教师模型的梯度作为对指定属性的响应。通过分层的梯度筛选，选取与指定属性内容相关的生成层，并定位到该层特征编码前 K 大的通道。所获取的编码通道与待编辑的属性呈现紧密联系。我们的方法利用少量的正样本快速地实现对属性的编码通道的定位。最后，提出了单通道和多通道的编辑方法，实现了基于内容区域和语义分类属性的人脸编辑。

7.2 未来研究方向

本文面向人脸编辑任务进行研究，并围绕图像特征的编码学习、特征解码生成和特征编辑展开。本文采取递进式的研究思路，是一种纵向的研究过程。图像特征是研究的核心内容基础。其中，人脸特征编码众多图像任务的基础。特征解码是基于特征编码网络，并学习过程中相互影响。图像特征编辑是在图像特征编码和解码的研究上，进一步探索特征编码空间和图像语义空间的关联。因此，缺少对于图像特征的横向研究内容的讨论。目前，关于图像特征编码的学习，解码和编辑方面的研究正快速地发展，涌现了许多有价值的研究方向。本文依据现有的工作，对所研究的关于图像特征的问题进行了部分展望。此外，由于人脸对齐任务相对简单，技术已经较为成熟。人脸对齐任务的拓展大都面向局部点云数据的对齐，这一大大领域超出了本文人脸编辑的探讨范围。

1. 图像或点云对齐的研究方向。本文利用三维点云数据实现图像配准。对于二维平面内的人脸图像可以围绕人脸关键点展开对齐研究。然而，对于表情丰富的柔性人脸，二维的关键点难以表示空间上的变换关系。为实现精细度更高的对齐，我们将目光聚焦于三维的点云数据的对齐研究。点云对齐研究在难度上远超二维人脸对齐的研究。目前，点云对齐的研究主要三个主要的困难及研究方向。第一，局部点云对齐大都需要进行点匹配的过程。借助深度学习的特征匹配研究 [23]，如何利用周围点的特征实现点匹配是研究热点问题。第二，点云对齐的效率。不同于图像的二维矩阵数据表示，三维点云在数据表示上，空间上是稀疏的，在点的数量上是庞大的。高效地处理点云数据实现快速的局部点云对齐，是限制其应用范围的一大挑战。
2. 图像特征编码学习方法的研究方向。对比从标注图像数据学习的特征编码，从无标注图像数据中学习的特征编码在图像语义内容的表示上存在不足。为了学习更好的特征编码，相关的研究工作不再局限在手工标注或无标注的图像本身，而将目标转向互联网上海量的图像文本数据。由于图像文本数据对容易获取，CLIP[13] 利用其训练了大规模的图像编码模型，并在图像分类和目标检测等图像任务中验证了该方法在学习图像特征编码的优越性。CLIP 应用对比学习方法，既解决了标注图像数据繁琐的问题，也解决无标注图像数据下学习特征编码困难的问题。

3. 特征解码生成图像的研究方向。利用条件生成对抗网络进行特征解码，实现图像生成的工作已经被用于各种图像翻译任务中。对于上述方法，特征解码网络可以看作为由低分辨率到高分辨率图像的上采样卷积过程。最近，扩散模型（Diffusion Model）[16] 作为新的图像生成模型被提出，并在生成图像的质量上媲美生成对抗网络模型。隐扩散模型（Latent Diffusion Model）[100] 实现了在特征编码空间上的扩散，并稳定地生成高清图像。然而，该模型对特征解码的生成过程缺少语义内容上的研究。对于隐扩散模型的特征编码的解码过程，有待进一步发掘特征编码与生成语义内容的联系。
4. 基于特征的图像编辑研究方向。通过编辑特征实现图像编辑是本文研究的内容。然而，目前主流的实现图像编辑的方法都是采用输入提示词的方法，如 Stable Diffusion Model[101] 等。由于提示词大都为描述图像中的粗糙语义信息，基于提示词的方法无法实现对图像的细粒度修改和连续编辑。本文提出的图像特征的定位编辑方法可以与提示词方法进行结合，实现从输入层的提示词，中间层的图像特征编码，到输出层高清图像这三个过程的全方位可控编辑。同时，图像特征编辑技术也被用于多种图像应用中，如手绘草图生成 [102]，图像修复和补全 [103] 等。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 : 770 – 778.

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 3431 – 3440.

[3] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[4] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2009 : 248 – 255.

[5] LIN T-Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C] // European conference on computer vision. 2014 : 740 – 755.

[6] ZHAI X, OLIVER A, KOLESNIKOV A, et al. S4l: Self-supervised semi-supervised learning[C] // Proceedings of the IEEE conference on computer vision. 2019 : 1476 – 1485.

[7] WANG Y-X, GIRSHICK R, HEBERT M, et al. Low-shot learning from imaginary data[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018 : 7278 – 7286.

[8] 庄福振, 罗平, 何清, et al. 迁移学习研究进展 [J]. 软件学报, 2015, 26(1) : 26 – 39.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[11] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C] // International conference on machine learning. 2019 : 6105 – 6114.

[12] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(1) : 5485 – 5551.

[13] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C] // International conference on machine learning. 2021 : 8748 – 8763.

[14] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8) : 1798 – 1828.

[15] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[16] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33 : 6840 – 6851.

[17] JING Y, YANG Y, FENG Z, et al. Neural style transfer: A review[J]. IEEE transactions on visualization and computer graphics, 2019, 26(11) : 3365 – 3385.

[18] ANWAR S, KHAN S, BARNES N. A deep journey into super-resolution: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(3) : 1 – 34.

[19] WANG Y, SOLOMON J M. Deep closest point: Learning representations for point cloud registration[C] // Proceedings of the IEEE international conference on computer vision. 2019 : 3523 – 3532.

[20] WU Z, SONG S, KHOSLA A, et al. 3D ShapeNets: A deep representation for volumetric shapes[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 1912 – 1920.

[21] YUAN W, ECKART B, KIM K, et al. DeepGMR: learning latent Gaussian mixture models for registration[C] // European conference on computer vision. 2020 : 733 – 750.

[22] HANDA A, WHELAN T, MCDONALD J, et al. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM[C] // IEEE international conference on robotics and automation. 2014 : 1524 – 1531.

[23] CHOY C, DONG W, KOLTUN V. Deep global registration[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2020 : 2514 – 2523.

[24] OLAH C, MORDVINTSEV A, SCHUBERT L. Feature visualization[J]. Distill, 2017, 2(11) : e7.

[25] WANG R, CHEN J, YU G, et al. Attribute-specific control units in stylegan for fine-grained image manipulation[C] // Proceedings of the 29th ACM international conference on multimedia. 2021 : 926 – 934.

[26] LING H, KREIS K, LI D, et al. Editgan: High-precision semantic image editing[J]. Advances in Neural Information Processing Systems, 2021, 34 : 16331 – 16345.

[27] SHEN Y, YANG C, TANG X, et al. InterfaceGAN: Interpreting the disentangled face representation learned by gans[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

[28] TEWARI A, ELGHARIB M, BHARAJ G, et al. Stylerig: rigging stylegan for 3d control over portrait images[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2020 : 6142 – 6151.

[29] MISRA I, MAATEN L V D. Self-supervised learning of pretext-invariant representations[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2020 : 6707 – 6717.

[30] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[31] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles[J]. Lecture Notes in Computer Science, 2016 : 69 – 84.

[32] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles[C] // European conference on computer vision. 2016 : 69 – 84.

[33] SERMANET P, LYNCH C, CHEBOTAR Y, et al. Time-contrastive networks: Self-supervised learning from video[C] // IEEE international conference on robotics and automation (ICRA). 2018 : 1134 – 1141.

[34] TSAI Y-H H, WU Y, SALAKHUTDINOV R, et al. Self-supervised learning from a multi-view perspective[J]. arXiv preprint arXiv:2006.05576, 2020.

[35] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization[C] // European conference on computer vision. 2016 : 649 – 666.

[36] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised representation learning by predicting image rotations[J]. arXiv preprint arXiv:1803.07728, 2018.

[37] PARK T, EFROS A A, ZHANG R, et al. Contrastive learning for unpaired image-to-image translation[C] // Computer Vision–ECCV 2020. 2020 : 319 – 345.

[38] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[39] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2020 : 8110 – 8119.

[40] HÄRKÖNEN E, HERTZMANN A, LEHTINEN J, et al. Ganspace: discovering interpretable gan controls[J]. Advances in neural information processing systems, 2020, 33 : 9841 – 9850.

[41] ABDAL R, ZHU P, MITRA N J, et al. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows[J]. ACM Transactions on Graphics (TOG), 2021, 40(3) : 1 – 21.

[42] JAISWAL A, BABU A R, ZADEH M Z, et al. A survey on contrastive self-supervised learning[J]. Technologies, 2020, 9(1): 2.

[43] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[44] WU Z, XIONG Y, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733 – 3742.

[45] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2020: 9729 – 9738.

[46] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C] // International conference on machine learning. 2020: 1597 – 1607.

[47] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, et al. Learning deep representations by mutual information estimation and maximization[J]. arXiv preprint arXiv:1808.06670, 2018.

[48] TIAN Y, KRISHNAN D, ISOLA P. Contrastive multiview coding[C] // European conference on computer vision. 2020: 776 – 794.

[49] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[50] LI Z, USMAN M, TAO R, et al. A systematic survey of regularization and normalization in GANs[J]. ACM Computing Surveys, 2022.

[51] CHAUDHARI S, MITHAL V, POLATKAN G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology, 2021, 12(5): 1 – 32.

[52] GONZALEZ R C. Digital image processing[M]. [S.l.]: Pearson education india, 2009.

[53] BISHOP C M, OTHERS. Neural networks for pattern recognition[M]. [S.l.]: Oxford university press, 1995.

[54] STRUDEL R, GARCIA R, LAPTEV I, et al. Segmenter: Transformer for semantic segmentation[C] // Proceedings of the IEEE international conference on computer vision. 2021: 7262–7272.

[55] CHANG H, ZHANG H, JIANG L, et al. Maskgit: Masked generative image transformer[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2022: 11315–11325.

[56] ZHOU Z, SHIN J, ZHANG L, et al. Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4761–4772.

[57] PATASHNIK O, WU Z, SHECHTMAN E, et al. StyleCLIP: text-driven manipulation of stylegan imagery[C] // Proceedings of international conference on computer vision. 2021: 2085–2094.

[58] KAJI S, KIDA S. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging[J]. Radiological physics and technology, 2019, 12: 235–248.

[59] ISOLA P, ZHU J-Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125–1134.

[60] MINAEE S, BOYKOV Y Y, PORIKLI F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021.

[61] SHEN W, LIU R. Learning residual images for face attribute manipulation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4030–4038.

[62] WU Z, LISCHINSKI D, SINTERFACEGANHECHTMAN E. Stylespace analysis: Disentangled controls for StyleGAN image generation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2021 : 12863 – 12872.

[63] TOV O, ALALUF Y, NITZAN Y, et al. Designing an encoder for stylegan image manipulation[J]. ACM Transactions on Graphics (TOG), 2021, 40(4) : 1 – 14.

[64] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2021.

[65] WU W, QIAN C, YANG S, et al. Look at boundary: a boundary-aware face alignment algorithm[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018 : 2129 – 2138.

[66] DAI J S. Euler–Rodrigues formula variations, quaternion conjugation and intrinsic connections[J]. Mechanism and Machine Theory, 2015, 92 : 144 – 152.

[67] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C] // IEEE Symposium on Security and Privacy (sp). 2017 : 39 – 57.

[68] PAN L, CHEN X, CAI Z, et al. Variational Relational Point Completion Network[J]. arXiv preprint arXiv:2104.10154, 2021.

[69] ZENG A, SONG S, NIESSNER M, et al. 3Dmatch: learning local geometric descriptors from rgb-d reconstructions[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 1802 – 1811.

[70] DENG H, BIRDAL T, ILIC S. PPFNET: global context aware local features for robust 3d point matching[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018 : 195 – 205.

[71] BESL P J, MCKAY H D. A method for registration of 3D shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1992, 14(2) : 239 – 256.

[72] ZHOU Q-Y, PARK J, KOLTUN V. Fast global registration[C] // European conference on computer vision. 2016 : 766 – 782.

[73] LI J, ZHANG C, XU Z, et al. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration[C] // European conference on computer vision. 2020.

[74] FU K, LIU S, LUO X, et al. Robust point cloud registration framework based on deep graph matching[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2021 : 8893 – 8902.

[75] ECKART B, KIM K, KAUTZ J. HGMR: hierarchical gaussian mixtures for adaptive 3D registration[C] // European conference on computer vision. 2018 : 705 – 721.

[76] AOKI Y, GOFORTH H, SRIVATSAN R A, et al. Pointnetlk: robust & efficient point cloud registration using pointnet[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2019 : 7163 – 7172.

[77] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6) : 381 – 395.

[78] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[79] BULAT A, TZIMIROPOULOS G. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)[C] // Proceedings of the IEEE international conference on computer vision. 2017 : 1021 – 1030.

[80] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60 : 91 – 110.

[81] NGUYEN A, YOSINSKI J, CLUNE J. Understanding neural networks via feature visualization: a survey[J]. Explainable AI: interpreting, explaining and visualizing deep learning, 2019 : 55 – 76.

[82] YAN Y, HAO H, XU B, et al. Image clustering via deep embedded dimensionality reduction and probability-based triplet loss[J]. IEEE Transactions on Image Processing, 2020, 29 : 5652 – 5661.

[83] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C] // International conference on machine learning. 2016 : 478 – 487.

[84] GUO X, GAO L, LIU X, et al. Improved deep embedded clustering with local structure preservation[C] // Proceedings of the 26th international joint conference on artificial intelligence. 2017 : 1753 – 1759.

[85] GHASEDI DIZAJI K, HERANDI A, DENG C, et al. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization[C] // Proceedings of the IEEE international conference on computer vision. 2017 : 5736 – 5745.

[86] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11) : 2278 – 2324.

[87] WOLF L, HASSNER T, MAOZ I. Face recognition in unconstrained videos with matched background similarity[C] // Proc. IEEE Conf. Comput. Vis. Pattern Recog.. 2011.

[88] Sim T, Baker S, Bsat M. The CMU Pose, Illumination, and Expression (PIE) database[C] // Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition. 2002 : 53 – 58.

[89] ZHANG W, ZHAO D, WANG X. Agglomerative clustering via maximum incremental path integral[J]. Pattern Recognition, 2013, 46(11) : 3056 – 3065.

[90] LI F, QIAO H, ZHANG B. Discriminatively boosted image clustering with fully convolutional auto-encoders[J]. Pattern Recognition, 2018, 83 : 161 – 173.

[91] BÄCKLUND H, HEDBLOM A, NEIJMAN N. A density-based spatial clustering of application with noise[J]. Data Mining TNM033, 2011, 33 : 11 – 30.

[92] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2019 : 4401 – 4410.

[93] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C] // Proceedings of the IEEE international conference on computer vision. 2017 : 1501 – 1510.

[94] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2020.

[95] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[96] LIU Z, LUO P, WANG X, et al. Deep Learning Face Attributes in the Wild[C] // Proceedings of international conference on computer vision. 2015.

[97] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[98] YU C, WANG J, PENG C, et al. BiseNet: bilateral segmentation network for real-time semantic segmentation[C] // European conference on computer vision. 2018 : 325 – 341.

[99] KHODADADEH S, GHADAR S, MOTIIAN S, et al. Latent to latent: A learned mapper for identity preserving editing of multiple face attributes in stylegan-generated images[C] // Proceedings of the IEEE winter conference on applications of computer vision. 2022 : 3184 – 3192.

[100] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2022 : 10684 – 10695.

[101] HO J, SALIMANS T. Classifier-free diffusion guidance[J]. arXiv preprint arXiv:2207.12598, 2022.

[102] 陈健, 白琮, 马青, et al. 面向细粒度草图检索的对抗训练三元组网络 [J]. 软件学报, 2020, 31(7) : 1933 – 1942.

[103] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 : 2536 – 2544.

附录 A　符号与函数说明

A.1　基础符号声明

本文使用形如 $\boldsymbol{w}$ 的粗体小写字母表示向量，使用形如 w 的小写字母表示标量，使用形如 $\boldsymbol{W}$ 的粗体大写字母表示矩阵，使用形如 W 的大写字母表示模型或函数。使用形如 $\mathcal{A}$ 的花题大写字母表示集合。

A.2　基础函数说明

下文将对本文所使用的基础函数进行说明。

1. **Sigmoid 函数** (Sigmoid(·))。对于任意输入 $x \in \mathcal{R}$，Sigmoid(x) 定义为：

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}. \tag{A-1}$$

对于任意向量形式的输入 $\boldsymbol{x} = [x_1, x_2, ..., x_m] \in \mathcal{R}^m$，Sigmoid 函数作用于 $\boldsymbol{x}$ 的每一维，并输出对应维度的结果向量，表示为

$$\text{Sigmoid}(\boldsymbol{x}) = [\text{Sigmoid}(x_1), \text{Sigmoid}(x_2), ..., \text{Sigmoid}(x_m)]. \tag{A-2}$$

2. **Softmax 函数** (Softmax(·))。对于任意输入向量 $\boldsymbol{x} = [x_1, x_2, ..., x_m] \in \mathcal{R}^m$，Softmax 函数对 $\boldsymbol{x}$ 中的每一维进行指数归一化并保留向量形式，即对于任意 $1 \le i \le m$，均有

$$\text{Softmax}(\boldsymbol{x}) = \frac{e^{x_i}}{\sum_{j=1}^{m} e^{x_j}}. \tag{A-3}$$

3. **球面线性插值函数** (Slerp(·))。对于向量 p_0 和 p_1，以 p_0 为起点，p_1 为终点所构成的球面中，在插值点 t 的球面插值定义为：

$$\text{Slerp}(p_0, p_1; t) = \frac{\sin[(1-t)\Omega]}{\sin(\Omega)} p_0 + \frac{\sin[t\Omega]}{\sin\Omega} p_1. \tag{A-4}$$

其中 $\cos(\Omega) = \frac{p_0 \cdot p_1}{\|p_0\| \|p_1\|}$，$\Omega$ 为 p_0 和 p_1 的夹角。

A.3 概率分布

本节主要介绍本文中所涉及到的概率分布。对于每种分布，我们列举了概率密度函数及其重要参数。

1. **单元高斯分布**。对于单变量 $x \in (-\infty, \infty)$，单元高斯分布的参数为均值 $\mu \in (-\infty, \infty)$ 和标准差 σ，则该参数下高斯分布的概率密度为：

$$\begin{aligned} p\left(x \mid \mu, \sigma^2\right) = \mathcal{N}\left(x \mid \mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; \\ \mathbb{E}[x] &= \mu; \\ \operatorname{var}[x] &= \sigma^2. \end{aligned} \tag{A-5}$$

2. **多元高斯分布**。对于 d 维向量 $\boldsymbol{x}$，多元高斯分布的参数维 d 维的均值向量 $\boldsymbol{\mu}$ 和 $d \times d$ 的对称正定协方差矩阵 $\boldsymbol{\Sigma}$:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \end{aligned} \tag{A-6}$$

其中期望 $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$ 和方差 $\operatorname{var}[\boldsymbol{x}] = \boldsymbol{\sigma}^2$，$\det(\boldsymbol{\Sigma})$ 为计算对称正定协方差矩阵的行列式。

3. **学生 t 分布**。学生 t 分布中，关于随机变量 X 服从自由度为 p 的概率密度函数定义为：

$$f_X(t) = \frac{\Gamma\left(\frac{\mathrm{p}+1}{2}\right)}{\Gamma\left(\frac{\mathrm{p}}{2}\right)} \frac{1}{\sqrt{\mathrm{p}\pi}} \left(1 + \mathrm{t}^2/\mathrm{p}\right)^{-\frac{\mathrm{p}+1}{2}}, -\infty < \mathrm{t} < \infty, \tag{A-7}$$

其中 $\Gamma(a)$ 为 Gamma 函数表示为：

$$\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} \mathrm{dt}. \tag{A-8}$$

学生 t 分布的期望和方差为：

$$\begin{aligned} \mathbb{E}\left(\mathrm{X_p}\right) &= 0, \mathrm{p} > 1; \\ \operatorname{Var}\left(\mathrm{X_p}\right) &= \frac{\mathrm{p}}{\mathrm{p}-2}, \mathrm{p} > 2. \end{aligned} \tag{A-9}$$

当 $p = 1$ 时，学生 t 分布的期望不存在；当 $p \leq 2$ 时，t 学生 t 分布的方差不存在。学生 t 分布也可以由高斯分布和卡方分布变换得到。设由两个独立的随机变量 U 和 V，且变量 U 服从标准正态分布 $N(0,1)$，变量 V 服从自由度为 p 的卡方分布，则 $U/\sqrt{V/p}$ 服从一个自由度为 p 的学生 t 分布。

致　　谢

我在南京大学度过了五年难忘的研究生时光。这些年间中所遇到了人和事，使我收益良多。我内心满怀感激，相信这必将成为我人生中难以忘怀的珍贵记忆。此刻，我由衷地感谢在研究生期间给予我关心，支持和帮助的老师和同学。

首先，我要感谢我的导师申富饶教授。在科研方面，申老师治学严谨，坚持实事求是的原则，鼓励从问题出发的研究，并善于引导我从不同的角度去解决所面临的问题。同时，申老师在我研究期间给予我极大的自由度，支持我感兴趣的研究并尊重我的意愿。在科研实验和论文写作方面，申老师也不厌其烦地从每周与我展开细致地讨论，逐渐培养了我掌握独自进行科研的能力。在生活方面，申老师时刻关注我们研究生的心理健康问题，并随手拈来各种正反面的案例来缓解我们的压力。他教育我们不仅关注科研问题，也要注意身体心理健康建设。申老师常以自身为例，介绍其在生活和工作上的自我调节过程。再次感谢申老师的辛苦付出，使得我们在生活和科研中找到不足并取得进步。

其次，我要感谢电子科学与工程学院的赵健教授。赵健老师与申富饶老师展开了跨学科领域的深度交流，这大大拓展了我们的研究视野。赵健老师结合海外留学的经历，在英文写作方面给予我细致地指导。赵健老师曾多次分享论文写作经验，并对我浅陋的写作内容进行逐字逐句的检查并提出了各种有价值的改进意见，有效提高了我的论文写作能力。此外，赵老师也十分感兴趣我们的研究内容，从个人视角出发，在组会上提出了各种宝贵的意见，启发了我新的研究反向。

接着，我要感谢陪伴我度过研究生生涯的各位同学和伙伴。第一，感谢早已经毕业但和我同期进组的 2018 届硕士生同学，他们在科研和生活上陪伴我三年的博士时光。融洽的实验室氛围为我带来了许多欢乐和温暖。第二，感谢师兄师姐在刚我入实验室所给予的指导和关怀。第三，感谢师弟师妹的尊重和理解。随着博士生涯的前进，我也从刚入实验室的新人成长为最有资历的师兄。本着实验室的融洽交流的氛围，我也试着尽我所能在各方面帮助师弟师妹。但有时也存在好心办坏事，反向帮助了他们。但不曾收到相关的责备，感谢他们

的理解与包容。

最后，我要感谢我的家人。父亲母亲不遗余力且无条件地相信和支持我，为我提供了最坚强的后盾。我的姐姐也时常关注我的科研和生活，并以过来人的经验为我保驾护航。他们的支持使我面对在研究生涯的困难时无所畏惧，安心地完成学业。

博士生涯即将接近尾声，新的征程即将开启。但博士生涯不是终点，我相信读博期间的收获会伴随我的一生。

简历与科研成果

基本信息

严元杰，男，汉族，1996 年 5 月出生，浙江省衢州人。

教育背景

2018 年 9 月 — 2023 年 6 月 南京大学计算机科学与技术系 博士
2014 年 9 月 — 2018 年 6 月 吉林大学计算机科学与技术系 本科

攻读博士学位期间完成的学术成果

1. **Yuanjie Yan**, Hongyan Hao, Baile Xu, Jian Zhao, Furao Shen, “Image clustering via deep embedded dimensionality reduction and probability-based triplet loss”. IEEE Transactions on Image Processing, 2020, 29: 5652-5661.
2. **Yuanjie Yan**, Junyi An, Jian Zhao, Furao Shen. “Hybrid optimization with unconstrained variables on partial point cloud registration”[J]. Pattern Recognition, 2023, 136: 109267.
3. **Yuanjie Yan**, Suorong Yang, Yan Wang, Jian Zhao, Furao Shen. ”Review neural networks about image transformation based on IGC learning framework with annotated information[J]”. arXiv preprint arXiv:2206.10155, 2022.
4. **Yan, Yuanjie**, Jian Zhao and Furao Shen. “Attribute-specific manipulation based on layer-wise channels.”.
5. **Yuanjie Yan**, Yuxuan Bu, Furao Shen and Jian Zhao. ”Improving the transferability of adversarial examples with separated positive and negative Perturbations”.

攻读博士学位期间参与的科研课题

1. 国家自然科学基金“ 基于深度感知增强式联想记忆神经网络的信息融合系统研究”（课题年限 2019 年 1 月 — 2022 年 12 月），负责图像感知方面的研究。

2. 科技部重大项目" 基于神经可塑性的脉冲网络高效学习机制与类脑智能系统"，负责网络学习机制的研究。

攻读博士期间参与的发明专利

1. 葛轶洲，**严元杰**，卜宇轩，周青，赵健，申富饶。基于分离正负扰动生成对抗图像机器识别的方法。专利申请号：202010656484.0

攻读博士期间获得的比赛奖项

1. Yuanjie Yan and Juanyi An, ICCV Workshop 2021 Sensing, Understanding and Synthesizing Humans, 1^{st} Place Award in MVP Point Cloud Registration CHALLENGE.

《学位论文出版授权书》

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

_____年_____月_____日

论文题名	基于图像特征的人脸编辑研究				
研究生学号	DZ1833030	所在院系	计算机科学与技术系	学位年度	2023
论文级别	☐ 硕士　☐ 硕士专业学位 ☐ 博士　☐ 博士专业学位 （请在方框内画勾）				
作者 Email	1310135753@qq.com				
导师姓名	申富饶教授				

论文涉密情况：

☐ 不保密

☐ 保密，保密期（_____年_____月_____日　至　_____年_____月_____日）

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。