Journal of Image and Signal Processing 图像与信号处理, 2024, 13(4), 416-426
Published Online October 2024 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2024.134036

Hans汉斯

文档矢量化技术的研究进展与应用

王　彤，陆利坤

北京印刷学院信息工程学院，北京

收稿日期：2024年9月18日；录用日期：2024年10月11日；发布日期：2024年10月23日

摘　要

文档矢量化是一种将文档内容转化为数学向量表示的技术，一般来说就是将光栅图像或者栅格图像转换为矢量图像。通过矢量化，可以将文本数据转化为计算机可以理解和处理的形式，从而将文档资料通过计算机矢量化的格式(例如OFD，PDF等)完整地保存下来，为印刷过程中的文本处理、信息检索等领域提供了更多可能性。首先，介绍了文档矢量化的背景；其次，介绍了传统文档矢量化模型；然后，将传统方法到基于深度学习的方法进行了全面综述并对不同的方法进行了比较；最后，对文档矢量化的应用领域和发展进行探讨和展望。

关键词

文档矢量化，矢量图像，深度学习，自然语言处理

Research Progress and Application of Document Vectorization Technology

Tong Wang, Likun Lu

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Sep. 18th, 2024; accepted: Oct. 11th, 2024; published: Oct. 23rd, 2024

Abstract

Document vectorization is a technique that converts the content of a document into a mathematical vector representation, generally a raster image or raster image into a vector image. Through vectorization, the text data can be converted into a form that the computer can understand and process, so that the document data can be completely saved through the computer vectorized format (such as OFD, PDF, etc.), providing more possibilities for text processing, information retrieval and other fields in the printing process. Firstly, the background of document vectorization is introduced. Secondly, the traditional document vectorization model is briefly introduced. Then, the vectorization

文章引用: 王彤, 陆利坤. 文档矢量化技术的研究进展与应用[J]. 图像与信号处理, 2024, 13(4): 416-426.
DOI: 10.12677/jisp.2024.134036

and the key techniques of vectorization processing in recent years are introduced. Finally, the application fields and development of document vectorization are discussed and prospected.

Keywords

Document Vectorization, Vector Image, Deep Learning, Natural Language Processing

1. 引言

在信息爆炸的时代，海量的文本数据对数据处理技术提出了新的要求。如何有效地处理和利用这些文本数据成为一个重要的研究课题。文档矢量化技术通过将文档内容转化为数学向量，使计算机能够理解和处理这些数据。矢量图像具有内容描述与分辨率无关的特点，是一种更为紧凑的图像表示方法[1]。文档矢量化的基本步骤包括分词、统计词频、计算权重以及构建向量，这些步骤在实际应用中面临许多挑战，如处理多语言文档、文本的语义理解以及高效计算等。本文矢量图像的表示和图像矢量化进行了综述，描述了几种常见的矢量化数学模型和它们在矢量图像内容创建中的贡献。

2. 背景介绍

目前的文档矢量化方法也存在许多不足和缺点，主要包括以下几个方面：

1) 语义表达的不准确性：文本具有丰富的语义信息，但是这些信息的表达是复杂的、抽象的，因此存在着一定的不准确性。例如，同样一个词汇在不同的语境中可能有着不同的含义，这种多义性会影响文档矢量化的准确性。

2) 上下文信息的丢失：文本具有上下文依赖性，每个词汇在特定的语境中才能被准确理解。然而，在文档矢量化的过程中，由于无法考虑上下文信息，文本的一些重要信息可能会丢失或被错误处理。

3) 空间维度的限制：文本矢量化通常需要将文档表示为一个固定长度的数学向量，这导致了文本信息的丢失和压缩。例如，对于一个长篇小说或一部电影剧本，将其表示为一个固定长度的向量可能会丢失大量细节和情节信息。另外文档的格式容易丢失，而只注重与文档的内容。

4) 深层语义信息的把握困难：文本不仅有着琐碎的语言学信息，还涉及复杂的逻辑和语义推理，这更深层语义信息的把握是文档矢量化的难点。例如，读者通过阅读一段文字可以理解其中的情感、态度等因素，但这些深层语义信息难以在文档矢量化中被很好地表达。

5) 语义的主观性：语言是一种主观性强的交流工具，同样的文本对不同人有着不同的理解。这些主观的语义信息难以在文档矢量化的过程中被很好地处理。

综上所述，当前文档矢量化仍然存在着不足和缺点，这些问题的产生主要源于文本复杂的语言学、逻辑和语义特性，以及文本长度、主观性等方面的困难。为了更好地进行文档矢量化，需要利用最新的技术手段，持续提高文本矢量化的准确性和深度，例如结合自然语言处理、深度学习等技术，以获取更多细节、更准确的语义信息，实现文档信息的高效利用。

3. 文档矢量化模型概述

文档矢量化模型通过使用自然语言处理(NLP)和机器学习技术将文档或文本转换为向量，按照传统方

法、基于机器学习的文档矢量化方法和基于深度学习的文档矢量化方法进行分类，图 1 展示了文档矢量化的常见模型，各种模型的概述如下。

3.1. 传统文档矢量化模型

传统文档矢量化模型以 Bag-of-Words 模型和 TF-IDE 模型为代表。

1) Bag-of-Words 模型(BoW) [2]：词袋模型是文档矢量化最基本的方法之一，将文档表示为词的集合，并通过计算每个词的出现频率或其他权重值来构建文档向量。这种方法简单直观，易于实现，但忽略了词与词之间的顺序和语义关系。

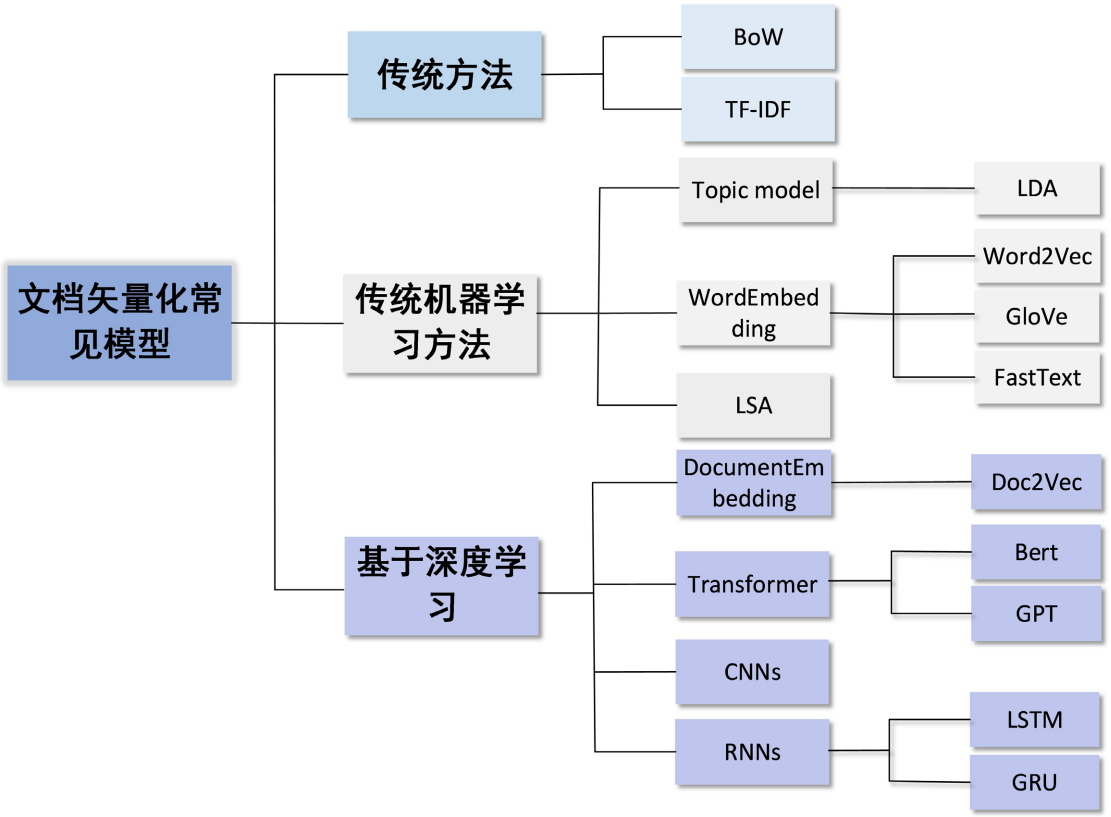

Figure 1. Common document vectorization models
图 1. 文档矢量化常见模型

2) TF-IDF 模型[3]：TF-IDF 又称为加权词袋模型，它在 BoW 的基础上，进一步考虑了单词在语料库中的重要性，计算的是该词在文档中的频率和在语料库中的逆文档频率之积。它将文档表示为词的权重向量。TF 表示词在文档中的频率，IDF 则表示逆文档频率，用于衡量词的重要性。TF-IDF 方法能够在考虑词频的同时，进一步获取词的重要性信息，对于信息检索问题表现较好。

3.2. 基于机器学习的文档矢量化模型

1) Word2Vec 模型[4]：Word2Vec 是一种基于神经网络的词嵌入模型，能够将词映射到一个低维空间，并保持词的语义关系。将单词映射为连续向量，通过无监督学习的方式得到单词向量。通过训练大规模语料库，Word2Vec 可以生成词的分布式表示。将 Word2Vec 应用于文档矢量化，可以通过计算文档中包含的词向量的平均值或加权平均值来构文档量。这种方法将语义信息引入了文档矢量化，能够更好地捕

捉词与词之间的语义关系。Word2Vec 模式以 CBOW 和 SkipGram 两个模型为代表。其中 x1, x2, ... , xN − 1, xN 表示一个文本中的 n-gram 向量，每个特征是词向量的平均值，用全部的 n-gram 去预测指定类别。

2) Topic model 模型：通过对文档的主题进行建模，得到文档的主题分布向量，例如 LDA、PLSA 等。LDA (Latent Dirichlet Allocation) [5]是一种主题模型，它将每个文档表示为一个概率分布向量，每个元素代表一个主题在该文档中出现的概率，文本集合中的每篇文档就可以表示成一个主题概率的分布集合。

3) GloVe 模型[6]：GloVe 模型是一种基于全局词汇统计信息的词向量表示方法，可以将每个单词表示为一个高维向量，也是一种将单词映射为连续向量的方法，与 Word2Vec 不同之处在于它采用了全局上下文信息，基于矩阵分解的思想充分利用了全局统计信息，能够更好地处理罕见词和多义词的语义表示。

4) FastText 模型[7]：FastText 模型是一种基于 Bag-of-Words 模型(词袋模型)的方法，通过将文本中的单词分解为字符级 n-gram 来捕捉单词内部的局部信息和词汇顺序信息，并通过这些信息进行文本分类，能够有效捕捉单词的前缀和后缀信息。

3.3. 基于深度学习的文档矢量化方法

3.3.1. Doc2vec 模型及其扩展

1) Doc2vec 模型[8]：将整个文档映射为连续向量，方法类似于 Word2Vec [9]，是 Word2Vec 的扩展，不仅可以学习到单词的向量表示，还可以学习到整个句子或文档的向量表示。

2) Sent2Vec 模型[10]：将整个句子映射为连续向量，采用了 Doc2vec 的原理，但是通过对训练过程中的单词进行微调，能够更好地捕捉句子或段落的语义。

3) Supervised embedding 模型：类似于 Word2Vec，应用于单词、短语、句子或文档表示的学习，但是采用有监督学习的方式，可以使用标注信息来指导向量学习。

4) Skip-thoughts 模型[11]：基于 Doc2Vec 的模型，通过学习句子之间的上下文关系，生成能够反映句子语义信息的向量表示，适用于类比推理等任务，属于无监督学习。

3.3.2. Bert 模型及其扩展

1) Bert 模型[12]：基于 Transformer 的预训练模型，可以对文本进行编码和解码，得到高质量的文本表示。

2) DistilBert 模型[13]：基于 Bert 模型的轻量级预训练模型，速度更快，但准确度相对较低。

3) BertScore 模型[14]：基于 Bert 模型的相似性度量工具，可以评估生成模型的输出和参考答案之间的相似性。

4) Albert 模型：基于轻量级 Transformer 的预训练模型，是在 Bert 模型的基础上进行改进和优化的，与 Bert 相比，Albert 具有更少的参数数量和更高的训练速度，其改进主要有参数共享和更高的训练速度的优势，可以在模型大小和准确度之间取得平衡。

3.3.3. 基于 Transformer 架构的模型

1) GPT-2 模型：基于 Transformer 的自回归语言模型，每一步生成一个单词，根据前面的词预测下一个词，擅长生成连贯的高质量文本。

2) T5 (Text-To-Text Transfer Transformer)模型：基于 Transformer 的通用文本生成模型，采用“文本到文本”的统一框架，所有任务都转换为文本生成任务。如翻译、摘要、问答等都视为输入文本生成输出文本的问题。

3) XLNet 模型：基于 Transformer-XL 的自回归语言模型，结合了 Bert 和自回归模型的优点，使用

“自回归并行化”的方法，对输入序列进行随机排列，训练模型预测序列中的空缺部分，可以捕捉更长的上下文信息。

以上列举了常见的文档矢量化方法，包括传统的统计方法、基于深度学习的模型以及传统机器学习算法等，这些方法在不同的场景和任务中都有其适用性和局限性，需要结合具体情况进行选择和优化。表 1 是各种文档矢量化模型的比较。

Table 1. Comparison of document vectorization models
表 1. 文档矢量化模型的比较

方法	优点	缺点
Bow	简单直观，计算效率高	无法捕捉语义关系和上下文信息
TF-IDF	比词袋模型(BoW)更能体现单词的重要性	仍然无法捕捉语义关系和上下文
Word2Vec	可以捕捉单词之间的语义和句法关系	需要大量训练数据，无法直接处理文档
GloVe	使用线性训练，计算速度快	没有考虑到上下文顺序信息
FastText	训练速度快，且能够有效捕捉到文本中的局部特征	无法捕捉到单词之间的顺序和语义关系
Topic model	可以发现文档的潜在主题，提供更丰富的语义信息	模型训练复杂，需要大量数据
Doc2Vec	可以直接表示文档，保留了语义信息	模型训练复杂，需要大量数据
BERT	可以捕捉复杂的上下文信息和语义关系，性能优异	模型规模庞大，计算资源要求高

4. 近年来矢量化以及矢量化处理关键技术

文档矢量化是自然语言处理(NLP)中的一个重要技术，它可以将文本数据转换为数字形式，以便计算机能够理解和处理，因此在文本分类、信息检索、文本摘要、文本相似度比较、机器翻译、语义分析等领域十分重要。

4.1. 基于深度学习的矢量化方法

传统位图(JPEG 或 PNG 格式图像)的分辨率较低、有难以进行几何操作以及文件大小过大的问题，因此将位图插图转换为具有可放缩性的矢量图像十分必要，然而传统的图像矢量化方法具有手工矢量化效率低下、难以处理复杂图形的缺点。深度矢量化模型与传统矢量化方法相比，无论是准确性还是效率方面都有显著的提高。

I-Chao Shen 等人[15]提出了一种基于深度生成模型的矢量化方法，用于将栅格图像转换为矢量图像。Shen 等人[16]提出的 ClipGen 模型通过生成对抗网络，自动学习插图特征并生成高质量的矢量图形。ClipGen 算法模型采用迭代生成的方式，使用一个全卷积编码器－解码器网络(FCN)和循环神经网络(RNN)解码器进行预测。ClipGen 网络结构图见图 2 所示。

Vage Egiazarian 等人针对手工矢量化效率低下和规模大、质量低的问题提出了一种基于深度学习的图纸矢量化方法[17]。该方法使用可学习的深度矢量化模型和新的基元优化方法，能够自动地将技术图纸转换为矢量格式。Ivan Puhachov 等人利用一种新颖的几何流(PolyVector Flow)，通过 PolyVector Flow 进行关键点驱动的线条绘制矢量化[18]。Mikhail 等人通过多矢量场对线条图进行矢量化，给定一个可能有噪声的灰度位图图像，通过使用该帧字段来提取绘图拓扑并使用计算的拓扑创建最终的矢量化[19]。Singh A K 等人提出了一个框架，用于识别与社交媒体上的热门话题相关的新闻文章，并进行多文档摘要。该框架使用了包括 TF-IDF、Word2Vec 和 Doc2Vec 的多种矢量化方法[20]。Ochilbek 提到了 Count (Binary)

vectorization 和 TF-IDF 两种矢量化方法[21]。其中，Count vectorization 方法将每个单词在句子中出现的情况用 0 或 1 表示，而 TF-IDF 方法则根据单词的重要性为其分配权重。

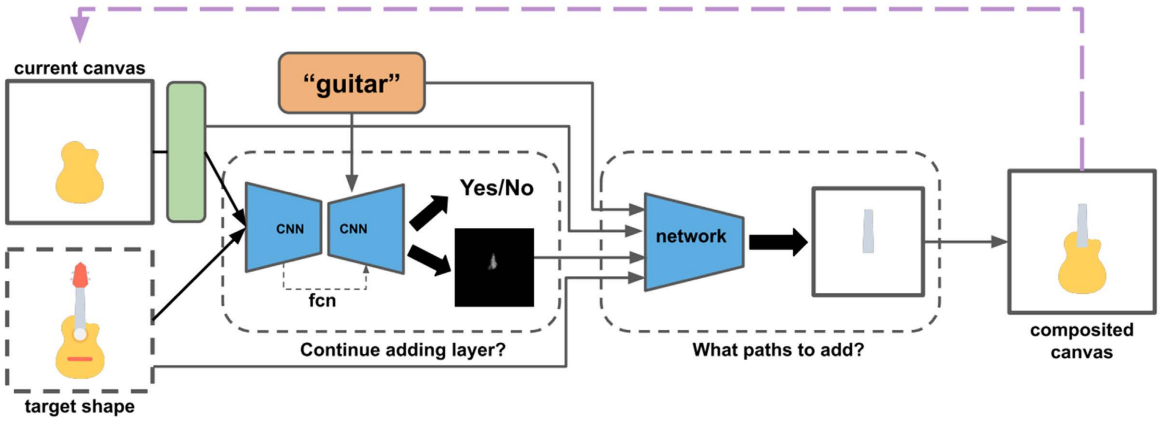

Figure 2. ClipGen model schematic diagram [16]
图 2. ClipGen 模型原理图[16]

Ayan 等人提出了矢量化(Vectorization)和栅格化(Rasterization)两个新的跨模态翻译预训练任务，用于自监督特征学习[22]。这些任务可以从未标记的手绘数据中学习强大的表示，并且在栅格和矢量两种不同的数据表示下都能够提高下游任务的性能。Diego 提出的基于曲线的基于图像边缘矢量化的描边算法[23]。通过简化内部连接，使得转换后的曲线段之间的连接更加平滑，从而减少了输出的线段数量，提高了算法的效率和准确性。Sagnik 等人提出了一个名为 DewarpNet 的端到端深度学习架构，用于单幅文档矫正[24]。DewarpNet 不需要使用多个图像或外部设备，可以在实时性的要求下进行文档图像去畸变。DewarpNet 包括两个网络：一个 3D 回归网络和一个 2D 回归网络。这两个网络可以同时预测文档的 3D 形状和 2D 畸变，从而实现文档图像的去畸变，DewarpNet 框架见图 3。

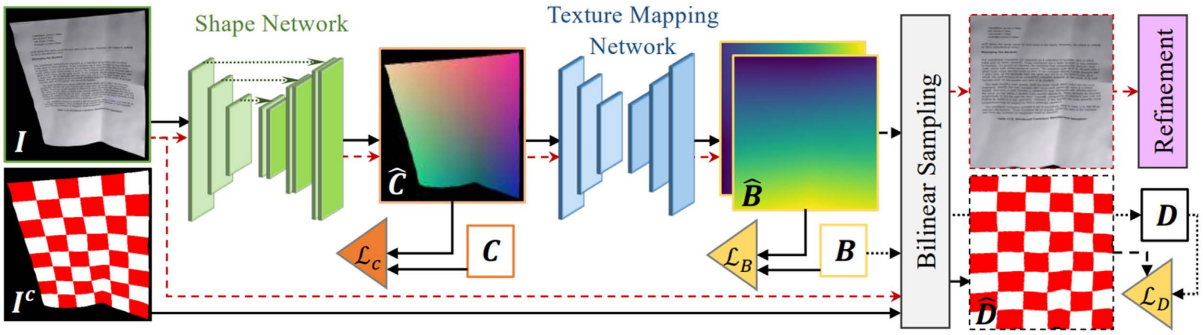

Figure 3. DewarpNet model schematic diagram [24]
图 3. DewarpNet 模型原理图[24]

Lee 等人提出了利用单词之间的网络信息来表示文档的关系特征的矢量化方法[25]。该方法使用单词的中心度量和单词在文档中与其他单词一起使用的次数来表示单词之间的关系。相比于传统的基于频率的方法，该方法能够更好地处理同义词和多义词的问题，从而提高文档向量化的准确性。Lee 等人在向量空间模型的文档排序方法的基础上进一步探索，提出了几种简化的向量空间模型实现方法，以降低计算复杂度并提高检索效果[26]。Mo 等人提出了一种基于笔画的矢量化方法，用于从图像中生成线描。通过优化笔画的位置和粗细来生成高质量的线描[27]。表 2 展示了基于深度学习的不同矢量化方法的优缺点。

Table 2. Vectorization method based on deep learning
表 2. 基于深度学习的矢量化方法

方法	时间	实现方法	优点	缺点
Doc2VeC	2017	将每个文档表示为单词嵌入的简单平均值	模型架构简单高效，可以快速进行训练和测试	对于长文档的表达能力较弱
DewarpNet	2019	使用端到端的深度学习架构	利用了文档纸张的 3D 形状表示，比仅使用 2D 表示的方法表现更好	无法捕捉到纸张褶皱这样的微小细节
Egiazarian *et al.*	2020	使用合成数据和真实数据训练的深度学习网络	初始向量和优化向量相结合能够，产生较为精确的向量表示	数据训练时需要大量人工标注的向量数据，对于包含复杂曲线的图像处理效果不佳
IVAN *et al.*	2021	构建多向量场后提取拓扑结构，再进行矢量化	能够较好处理角点处的线条交汇，能够处理多风格的线条图像	对有阴影的图像处理效果不佳
Stroke-to-Fil	2020	引入正则化阶段和内连接，简化方法处理线段细节	速度足够快，适用于大多数应用	内连接分类过程较为局部，无法处理非局部覆盖情况
Sketch2Vec	2021	利用手绘数据的双模态表，示进行自监督预训练	在手绘识别和检索任务上，预训练模型可以接近监督学习性能	方法主要关注特征表示的学习，没有及生成模型等其他自监督学习方法
Mo *et al.*	2021	学习从栅格图像空间到矢量图像空间的映射，从各种图像中生成线条绘图	可以处理各种类型的图像，训练和测试的速度较快	在极端复杂的图像中，模型可能无法捕捉所有细节
ClipGen	2022	通过 C-RNN 迭代生成新模型	方法支持多种应用	模块较为复杂，推理预测速度较慢

4.2. 面向多模态文档的矢量化方法

Naoto 等人引入了一种名为 FlexDM 的多任务学习方法模型，旨在解决矢量图形文档中的各种设计任务[28]。该方法利用基于 Transformer 的架构和掩码字段预测来生成完整的矢量图形文档。Zhao 等人提出了一种自动创建扩散曲线图像的矢量化方法[29]。该方法通过自动化流程、曲线几何形状优化、时间连贯性等多个方面进行了创新。传统的基于文本的搜索方法在文档中存在一些局限性，无法充分利用文档中的视觉和布局信息。为了解决这个问题，Abhinav 等人提出了一种名为“Monomer”的多模态融合方法，通过一次性片段检测(One-Shot Snippet Detection)来实现文档中基于视觉相似性的搜索[30]。此方法通过融合视觉、文本和空间模态的上下文信息，在目标文档中找到与查询片段相似的片段。

Chen 等人针对 LLVM 在自动矢量化过程中无法支持特定的 LCSSA (Loop-Closed SSA)循环的问题，提出了一种重构 PHI 节点的算法[31]。通过添加新的基本块，重构 PHI 节点的值，消除循环中外部使用对自动矢量化的影响，从而实现循环的自动矢量化。Li 等人提出了一个用于跨领域文档目标检测的基准套件，建立了一个包含不同类型的 PDF 文档数据集的基准套件，用于进行跨领域文档目标检测模型的训练和评估[32]。该套件包括页面图像、边界框注释、PDF 文件和从 PDF 文件中提取的渲染层等关键组件。引入了三个新颖的领域对齐模块：跨领域文档目标检测模型基于 Feature Pyramid Networks (FPN)目标检测器，并引入了三个新颖的领域对齐模块：Feature Pyramid Alignment (FPA)模块、Region Alignment (RA)模块和 Rendering Layer Alignment (RLA)模块。Li 等人提出了一种名为 SelfDoc 的多模态框架，用于处理文档理解任务[33]。该框架结合了语言和视觉模态，通过跨模态编码器、预训练阶段和模态自适应注意力模块来提取有意义的信息。Lin 等人提出了一种名为 BEDSR-Net 的深度学习网络，用于从单个文档图像

中去除阴影[34]。BEDSR-Net 原理图见图 4。Ma 等人提出了一种基于深度学习的方法，用于解决文档图像阴影去除问题，通过将文档的图像颜色和结构信息进行矢量化表示来提取文档的全局背景颜色[35]。该方法使用卷积神经网络(CNN)进行端到端的图像恢复，而不是传统的优化过程。Ma 等人提出了一种名为 LIVE 的层次化图像矢量化方法，可以将光栅图像转化为具有层次表示的矢量图形[36]。与之前的方法相比，LIVE 是无模型的，不需要形状基元标签，可以适用于各种领域的图像矢量化任务，引入了一种新的组件化路径初始化方法和一种新的损失函数，包括无符号距离引导的焦点损失函数(UDF Loss)和自交叉损失函数(Xing Loss)。这些方法可以改善从光栅图像生成矢量图形的质量，减少曲线交叉和形状失真。

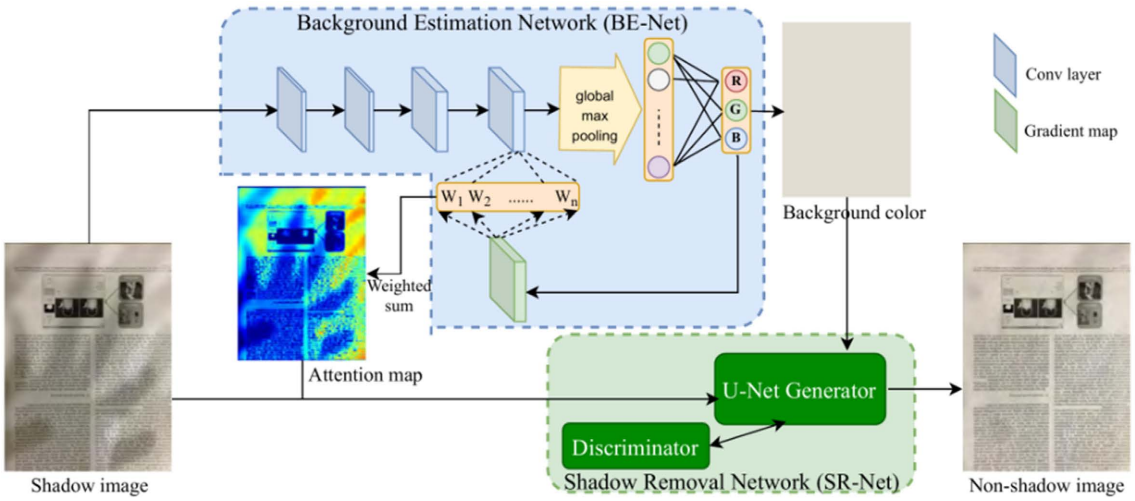

Figure 4. BEDSR-Net model schematic diagram [21]
图 4. BEDSR-Net 模型原理图[21]

Muhammad 等人使用大型语言模型(LLM)生成多个文本描述，提出了一种名为 I2MVFormer 的模型，利用 LLM 生成多个类别的多个文本描述(视图) [37]。Hoshyari 等人针对艺术家生成的半结构化栅格图像，提出了一种感知驱动的边界矢量化方法[38]。该方法通过利用准确性、连续性、简洁性和闭合性等感知线索，自动将离散的区域边界转化为分段平滑的曲线。Qi 等人提出了一种基于深度学习的弱监督方法，用于去除印刷和手写文本图像中的伪影[39]。该方法使用全卷积网络生成二进制分割掩模，指示每个像素中是否存在伪影。Bau 等人提出了一种利用生成对抗网络(GANs)的图像先验进行语义照片编辑的方法[40]。提出了一种训练网络来推断潜在向量 z 的方法。通过优化潜在向量 z，可以实现对图像的准确重建和编辑。展示了多种语义图像编辑任务中的有效性，包括合成新对象、删除不需要的对象特征及改变对象的外观等。Song 等人提出了一种用于建筑蓝图矢量化的新算法[41]。传统的建筑蓝图矢量化算法通常通过检测角点来开始，但对于具有细小内墙、小门框和长外墙高清蓝图的处理效果不佳。因此，本文提出了一种改进的生成对抗网络方法，该算法能够准确提取和表示建筑蓝图的复杂几何结构。与现有方法相比，该算法在定性和定量评估中展现出有希望的结果，并在标准矢量化指标上取得了显著的提升。

Othman 等人提出了一种新的深度图像生成范式，通过学习参数化的图层分解来生成图像[42]。与常用的卷积网络架构相比，这种分层分解的方法具有许多优势。Ding 等人提出了一种端到端的 PivotNet 框架来将地图进行矢量化[43]。PivotNet 通过将地图构建任务形式化为一个稀疏集合预测问题，并利用基于序列匹配的双边匹配损失，学习出精确而紧凑的矢量化表示，无需任何后处理。总体来说，PivotNet 通过枢轴点表示和动态序列匹配的方法，实现了从车载图像数据到地图矢量化表示的端到端学习。表 3 展示了面向多模态文档的不同矢量化方法的比较。

Table 3. Vectorization method for multimodal documents
表 3. 面向多模态文档的矢量化方法

方法	时间	实现方法	优点	缺点
SHAYAN *et al.*	2018	采用同时拟合样条曲线和角点检测的方法	生成的半结构化光栅图像，能够很好符合人类视觉感知	只适用于清晰的量化图像，对于抗锯齿或噪声数据的处理效果不佳
DeepErase	2019	使用弱监督学习训练 Artifacts 分割网络	视觉效果好	训练集和测试集分不存在偏差，测试集效果不如预期
Li *et al.*	2020	引入三个模块：特征金字塔对齐、区域对齐、渲染层对齐	通过三个模块有效缓解了域间差异问题	方法主要访问源域的标注数据，在标注数据受限的情况下难以大规模应用
BEDSER-Net	2020	引入背景估计模块，在合成图像上进行训练	相比其他文档阴影去除方法，在视觉效果和内容保留方面都取得了较好的效果	当文档完全处于阴影中或者有多个光源造成的复杂阴影时，该方法也可能失效
SelfDoc	2021	采用语义上有意义的组件作为输入以避免过细的上下文关联	需要较少的预训练数据就可以获得很好的下游任务表现	组件级输入可能会丢失一些词级信息
LIVE	2022	逐层将栅格图像转换为 SVG 格式，同时保持图像的拓扑结构	在相同的路径数量下，LIVE 能够取得更好地重构性能	分层操作的效率不如单次优化的方法
FlexDM	2023	基于 Transformer 的架构和掩码字段预测来生成完整的矢量图形文档	一种较为通用的方法来表示和解决多种设计任务	当输入文档元素较多时，模型性能会下降
MONOMER	2023	将任务作为一个一次性片段检测任务来解决	利用查询片段和目标文档的视觉、文本和空间模态信息，提高了片段检测性能	模型较为复杂，需要多个编码器模块和注意力机制，计算成本较高
I2MVFormer	2023	使用大型语言模型(LLM)生成每个类别的多个文本描述，以提供额外的监督信息	利用 LLM 生成的多视角文本提供了互补的信息，提高了零样本图像分类的性能	模型复杂度较高，训练时间较长

5. 总结与展望

本文回顾了文档矢量化的方法和近年来具有代表性的文档矢量化技术。文档矢量化作为将非结构化文本转换为结构化向量表示的关键技术，在自然语言处理和信息检索领域扮演着重要角色。该技术的核心思想是通过建模文档中的词语的词义和统计信息，将文档映射到一个联系的向量空间中，使语义相似的文档在该向量空间中彼此靠近。可以说，文档矢量化技术通过自动捕捉文档的语义，大大降低了文本处理的复杂度，为下游任务如文本聚类、相似度计算、主题建模等提供了强有力的支持。文档矢量化技术在不断进步，未来的向量表示在多模态信息融合、结合知识图谱的矢量化等方面存在发展空间。随着深度学习、多模态数据处理等技术的发展，文档矢量化将展现出更大的潜力，并在未来的研究和应用中得到进一步的发展和应用。

基金项目

北京市教育委员会出版学新兴交叉学科平台建设－数字喷墨印刷技术及多功能轮转胶印机关键技术研发平台(04190123001/003)；北京市数字教育研究重点课题(BDEC2022619027)；北京市高等教育学会 2023 年立项面上课题(课题编号：MS2023168)；北京印刷学院校级科研项目(20190122019, Ec202303, Ea202301, E6202405)；北京印刷学院学科建设和研究生教育专项(21090122012, 21090323009)；北京市自然科学基金资助项目(1212010)。

参考文献

[1] Tian, X. and Günther, T. (2024) A Survey of Smooth Vector Graphics: Recent Advances in Representation, Creation, Rasterization, and Image Vectorization. *IEEE Transactions on Visualization and Computer Graphics*, **30**, 1652-1671. https://doi.org/10.1109/tvcg.2022.3220575

[2] Le, Q.V. and Mikolov, T. (2014) Distributed Representations of Sentences and Documents. *The* 31*st International Conference on Machine Learning* (*ICML* 2014), Beijing, 21-26 June 2014, 1188-1196.

[3] Grootendorst, M. (2022) BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure.

[4] Tomas, M., Ilya, S., Kai, C., Greg, C., Jeffrey, D., *et al.* (2013) Distributed Representations of Words and Phrases and their Compositionality. *Conference on Neural Information Processing Systems*, Lake Tahoe, 5-10 December 2013, 3111-3119.

[5] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., *et al*. (2018) Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools and Applications*, **78**, 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[6] Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1532-1543. https://doi.org/10.3115/v1/d14-1162

[7] Armand, J., Edouard, G., Piotr, B., Tomas, M., *et al*. (2017) Bag of Tricks for Efficient Text Classification. *Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, 3-7 April 2017, 427-431.

[8] Qader, W.A., Ameen, M.M. and Ahmed, B.I. (2019) An Overview of Bag of Words: Importance, Implementation, Applications, and Challenges. 2019 *International Engineering Conference* (*IEC*), Erbil, Iraq, 23-25 June 2019, 200-204.

[9] Tomás, M., Kai, C., Greg, C., Jeffrey, D., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space. Computing Research Repository.

[10] Arora, S., Liang, Y.Y. and Ma, T.Y. (2017) A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *International Conference on Learning Representations*, Toulon, 24-26 April 2017, 1-16.

[11] Ryan, K., Yukun, Z., Ruslan, S., Richard, S.Z., Antonio, T., Raquel, U., Sanja, F., *et al*. (2015) Skip-Thought Vectors. *Advances in Neural Information Processing Systems* 28: *Annual Conference on Neural Information Processing Systems* 2015, Montreal, 7-12 December 2015, 3294-3302.

[12] Jacob, D., Kenton, L., Kristina, T., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics, 4171-4186.

[13] Sanh, V., Debut, L., Chaumond, J., Wolf, T., *et al*. (2019) DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. Obstetric Protocols for Labor Ward Management.

[14] Varsha, K., Felix, W., Kilian, Q.W., Yoav, A., *et al*. (2020) BERTScore: Evaluating Text Generation with BERT. *International Conference on Learning Representations*, Addis Ababa, 30 April 2020, 1904.

[15] Shen, I. and Chen, B. (2022) Clipgen: A Deep Generative Model for Clipart Vectorization and Synthesis. *IEEE Transactions on Visualization and Computer Graphics*, **28**, 4211-4224. https://doi.org/10.1109/tvcg.2021.3084944

[16] Shen, L.X., Shen, E., Tai, Z.W., Xu, Y.H., Dong, J.X. and Wang, J.M. (2022) Visual Data Analysis with Task-Based Recommendations. *Data Science and Engineering*, **7**, 354-369.

[17] Egiazarian, V., Voynov, O., Artemov, A., Volkhonskiy, D., Safin, A., Taktasheva, M., *et al*. (2020) Deep Vectorization of Technical Drawings. 16*th European Conference*, Glasgow, 23-28 August 2020, 582-598. https://doi.org/10.1007/978-3-030-58601-0_35

[18] Bessmeltsev, M. and Solomon, J. (2019) Vectorization of Line Drawings via Polyvector Fields. *ACM Transactions on Graphics*, **38**, 1-12. https://doi.org/10.1145/3202661

[19] Mikhail, B. and Justin, S. (2019) Vectorization of Line Drawings via Polyvector Fields. *ACM Transactions on Graphics*, **38**, Article No. 9.

[20] Singh, A.K. and Shashi, M. (2019) Vectorization of Text Documents for Identifying Unifiable News Articles. *International Journal of Advanced Computer Science and Applications*, **10**, 305-310. https://doi.org/10.14569/ijacsa.2019.0100742

[21] Rakhmanov, O. (2020) A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. *Procedia Computer Science*, **178**, 194-204. https://doi.org/10.1016/j.procs.2020.11.021

[22] Bhunia, A.K., *et al*. (2021) Vectorization and Rasterization: Self-Supervised Learning for Sketch and Handwriting. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 5668-5677.

[23] Nehab, D. (2020) Converting Stroked Primitives to Filled Primitives. *ACM Transactions on Graphics*, **39**, 137:1-137:17. https://doi.org/10.1145/3386569.3392392

[24] Das, S., *et al*. (2019) DewarpNet: Single-Image Document Unwarping with Stacked 3D and 2D Regression Networks. *IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 131-140.

[25] Lee, S.Y. (2019) Document Vectorization Method Using Network Information of Words. *PLOS ONE*, **14**, e0219389. https://doi.org/10.1371/journal.pone.0219389

[26] Lee, D.L., Chuang, H. and Seamons, K. (1997) Document Ranking and the Vector-Space Model. *IEEE Software*, **14**, 67-75. https://doi.org/10.1109/52.582976

[27] Chen, M.M. (2017) Efficient Vector Representation for Documents through Corruption. *International Conference on Learning Representations*, Toulon, 24-26 April 2017, 24-26.

[28] Mo, H., Simo-Serra, E., Gao, C., Zou, C. and Wang, R. (2021) General Virtual Sketching Framework for Vector Line Art. *ACM Transactions on Graphics*, **40**, 1-14. https://doi.org/10.1145/3450626.3459833

[29] Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M. and Yamaguchi, K. (2023) Towards Flexible Multi-Modal Document Models. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14287-14296. https://doi.org/10.1109/cvpr52729.2023.01373

[30] Zhao, S., Durand, F. and Zheng, C. (2018) Inverse Diffusion Curves Using Shape Optimization. *IEEE Transactions on Visualization and Computer Graphics*, **24**, 2153-2166. https://doi.org/10.1109/tvcg.2017.2721400

[31] Java, A., Deshmukh, S., Aggarwal, M., Jandial, S., Sarkar, M. and Krishnamurthy, B. (2023) One-Shot Doc Snippet Detection: Powering Search in Document Beyond Text. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 2-7 January 2023, 5426-5435. https://doi.org/10.1109/wacv56688.2023.00540

[32] Chen, M., Chai, Y. and Shang, J. (2021) LCSSA Optimization for Vectorization Recognition Rate Improvement. *Journal of Physics*: *Conference Series*, **1827**, Article ID: 012143. https://doi.org/10.1088/1742-6596/1827/1/012143

[33] Li, K., *et al*. (2020) Cross-Domain Document Object Detection: Benchmark Suite and Method. *Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 12912-12921.

[34] Li, P.Z., *et al*. (2021) SelfDoc: Self-Supervised Document Representation Learning. *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 19-25 June 2021, 5652-5660.

[35] Lin, Y., Chen, W. and Chuang, Y. (2020) Bedsr-Net: A Deep Shadow Removal Network from a Single Document Image. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 12902-12911. https://doi.org/10.1109/cvpr42600.2020.01292

[36] Ma, K., *et al*. (2018) DocUNet: Document Image Unwarping via A Stacked U-Net. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-22 June 2018, 4700-4709.

[37] Ma, X., *et al*. (2022) Towards Layer-Wise Image Vectorization. *Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 16293-16302.

[38] Naeem, M.F., *et al*. (2023) I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification. *CVPR* 2023, Vancouver, 17-24 June 2023, 15169-15179.

[39] Hoshyari, S., Dominici, E.A., Sheffer, A., Carr, N., Wang, Z., Ceylan, D., *et al*. (2018) Perception-Driven Semi-Structured Boundary Vectorization. *ACM Transactions on Graphics*, **37**, Article No. 118. https://doi.org/10.1145/3197517.3201312

[40] Qi, Y., Huang, W.R., Li, Q. and DeGange, J.L. (2020) Deeperase: Weakly Supervised Ink Artifact Removal in Document Text Images. 2020 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Snowmass Village, 1-5 March 2020, 3511-3519. https://doi.org/10.1109/wacv45572.2020.9093532

[41] Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J., *et al*. (2019) Semantic Photo Manipulation with a Generative Image Prior. *ACM Transactions on Graphics*, **38**, Article No. 59. https://doi.org/10.1145/3306346.3323023

[42] Song, W., Abyaneh, M.M., Shabani, M.A. and Furukawa, Y. (2023) Vectorizing Building Blueprints. 16*th Asian Conference on Computer Vision*, Macao, 4-8 December 2022, 142-157. https://doi.org/10.1007/978-3-031-26319-4_9

[43] Ding, W.J., Qiao, L.M., Qiu, X., *et al*. (2023) PivotNet: Vectorized Pivot Learning for End-to-End HD Map Construction. *IEEE International Conference on Computer Vision*, Paris, 1-6 October 2023, 3649-3659.