中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于深度学习的直播电商监管关键技术研究

作者姓名：孙志东

指导教师：李学庆 教授 山东大学软件学院

学位类别：工学博士

学科专业：软件理论

培养单位：山东大学软件学院

2022 年 6 月

Research On Key Technologies Of Living Stream Ecommerce Supervision Based On Deep Learning

A dissertation submitted to

Shandong University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

in Fluid Mechanics

By

Sun Zhidong

Supervisor: Professor Liu Xueqing

School of Software Engineering, Shandong University

June, 2022

山东大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

山东大学
学位论文授权使用声明

本人完全了解并同意遵守软件学院有关保存和使用学位论文的规定，即软件学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

本文是中国科学院大学学位论文模板 ucasthesis 的使用说明文档。主要内容为介绍 LaTeX 文档类 ucasthesis 的用法，以及如何使用 LaTeX 快速高效地撰写学位论文。

关键词：深度学习，直播电商，监管技术，视频理解

Abstract

This paper is a help documentation for the LaTeX class ucasthesis, which is a thesis template for the University of Chinese Academy of Sciences. The main content is about how to use the ucasthesis, as well as how to write thesis efficiently by using LaTeX.

Keywords: Deep Learning, Live streaming, Supervision technology, Video Understanding

目 录

图形列表

表格列表

符号列表

字符

Symbol	Description	Unit
R	the gas constant	$m^2 \cdot s^{-2} \cdot K^{-1}$
C_v	specific heat capacity at constant volume	$m^2 \cdot s^{-2} \cdot K^{-1}$
C_p	specific heat capacity at constant pressure	$m^2 \cdot s^{-2} \cdot K^{-1}$
E	specific total energy	$m^2 \cdot s^{-2}$
e	specific internal energy	$m^2 \cdot s^{-2}$
h_T	specific total enthalpy	$m^2 \cdot s^{-2}$
h	specific enthalpy	$m^2 \cdot s^{-2}$
k	thermal conductivity	$kg \cdot m \cdot s^{-3} \cdot K^{-1}$
S_{ij}	deviatoric stress tensor	$kg \cdot m^{-1} \cdot s^{-2}$
τ_{ij}	viscous stress tensor	$kg \cdot m^{-1} \cdot s^{-2}$
δ_{ij}	Kronecker tensor	1
I_{ij}	identity tensor	1

算子

Symbol	Description
Δ	difference
∇	gradient operator
$\delta^{\pm}$	upwind-biased interpolation scheme

缩写

CFD	Computational Fluid Dynamics
CFL	Courant-Friedrichs-Lewy
EOS	Equation of State

JWL	Jones-Wilkins-Lee
WENO	Weighted Essentially Non-oscillatory
ZND	Zel'dovich-von Neumann-Doering

第 1 章　引言

1.1　研究背景

考虑到许多同学可能缺乏 LaTeX 使用经验，ucasthesis 将 LaTeX 的复杂性高度封装，开放出简单的接口，以便轻易使用。同时，对用 LaTeX 撰写论文的一些主要难题，如制图、制表、文献索引等，进行了详细说明，并提供了相应的代码样本，理解了上述问题后，对于初学者而言，使用此模板撰写学位论文将不存在实质性的困难。所以，如果你是初学者，请不要直接放弃，因为同样为初学者的我，十分明白让 LaTeX 简单易用的重要性，而这正是 ucasthesis 所追求和体现的。

此中国科学院大学学位论文模板 ucasthesis 基于中科院数学与系统科学研究院吴凌云研究员的 CASthesis 模板发展而来。当前 ucasthesis 模板满足最新的中国科学院大学学位论文撰写要求和封面设定。兼顾操作系统：Windows，Linux，MacOS 和 LaTeX 编译引擎：pdflatex，xelatex，lualatex。支持中文书签、中文渲染、中文粗体显示、拷贝 PDF 中的文本到其他文本编辑器等特性。此外，对模板的文档结构进行了精心设计，撰写了编译脚本提高模板的易用性和使用效率。

ucasthesis 的目标在于简化学位论文的撰写，利用 LaTeX 格式与内容分离的特征，模板将格式设计好后，作者可只需关注论文内容。同时，ucasthesis 有着整洁一致的代码结构和扼要的注解，对文档的仔细阅读可为初学者提供一个学习 LaTeX 的窗口。此外，模板的架构十分注重通用性，事实上，ucasthesis 不仅是国科大学位论文模板，同时，通过少量修改即可成为使用 LaTeX 撰写中英文文章或书籍的通用模板，并为使用者的个性化设定提供了接口。

1.2　系统要求

ucasthesis 宏包可以在目前主流的 LaTeX 编译系统中使用，如 TeXLive 和 MiKTeX。因 CTeX 套装已停止维护，**不再建议使用**（请勿混淆 CTeX 套装与 ctex 宏包。CTeX 套装是集成了许多 LaTeX 组件的 LaTeX 编译系统。ctex 宏包如同 ucasthesis，是 LaTeX 命令集，其维护状态活跃，并被主流的 LaTeX 编译系统默认集成，是几乎所有 LaTeX 中文文档的核心架构）。推荐的 LaTeX 编译系统 和 LaTeX 文本编辑器 为

操作系统	LaTeX 编译系统	LaTeX 文本编辑器
Linux	TeXLive Full	Texmaker 或 Vim
MacOS	MacTeX Full	Texmaker 或 Texshop
Windows	TeXLive Full 或 MiKTeX	Texmaker

LaTeX 编译系统，如 TeXLive（MacTeX 为针对 MacOS 的 TeXLive)，用于提供编译环境，LaTeX 文本编辑器 (如 Texmaker) 用于编辑 TeX 源文件。请从各软件官网下载安装程序，勿使用不明程序源。**LaTeX 编译系统和 LaTeX 编辑器分别安装成功后，即完成了 LaTeX 的系统配置**，无需其他手动干预和配置。若系统原带有旧版的 LaTeX 编译系统并想安装新版，请**先卸载干净旧版再安装新版**。

1.3 问题反馈

请见 问题反馈

欢迎大家有效地反馈模板不足之处，一起不断改进模板。希望大家向同事积极推广 LaTeX，一起更高效地做科研。

1.4 模板下载

Github/ucasthesis: https://github.com/mohuangrui/ucasthesis

第 2 章　LaTeX 使用说明

为方便使用及更好地展示 LaTeX 排版的优秀特性，ucasthesis 的框架和文件体系进行了细致地处理，尽可能地对各个功能和板块进行了模块化和封装，对于初学者来说，众多的文件目录也许一开始让人觉得有些无所适从，但阅读完下面的使用说明后，会发现原来使用思路是简单而清晰的，而且，当对 LaTeX 有一定的认识和了解后，会发现其相对 Word 类排版系统极具吸引力的优秀特性。所以，如果是初学者，请不要退缩，请稍加尝试和坚持，以领略到 LaTeX 的非凡魅力，并可以通过阅读相关资料如 LaTeX Wikibook (Wikibook, 2014) 来完善自己的使用知识。

2.1　先试试效果

1. 安装软件：根据所用操作系统和章节 1.2中的信息安装 LaTeX 编译环境。

2. 获取模板：下载 ucasthesis 模板并解压。ucasthesis 模板不仅提供了相应的类文件，同时也提供了包括参考文献等在内的完成学位论文的一切要素，所以，下载时，推荐下载整个 ucasthesis 文件夹，而不是单独的文档类。

3. 编译模板：

(a) Windows：双击运行 artratex.bat 脚本。

(b) Linux 或 MacOS: terminal -> chmod +x ./artratex.sh -> ./artratex.sh xa

(c) 任意系统：都可使用 LaTeX 编辑器打开 Thesis.tex 文件并选择 xelatex 编译引擎进行编译。

4. 错误处理：若编译中遇到了问题，请先查看“常见问题”（章节 2.4）。

编译完成即可获得本 PDF 说明文档。而这也完成了学习使用 ucasthesis 撰写论文的一半进程。什么？这就学成一半了，这么简单???，是的，就这么简单！

2.2　文档目录简介

2.2.1　Thesis.tex

Thesis.tex 为主文档，其设计和规划了论文的整体框架，通过对其的阅读可以了解整个论文框架的搭建。

2.2.2 编译脚本

- Windows：双击 Dos 脚本 artratex.bat 可得全编译后的 PDF 文档，其存在是为了帮助不了解 LaTeX 编译过程的初学者跨过编译这第一道坎，请勿通过邮件传播和接收此脚本，以防范 Dos 脚本的潜在风险。
- Linux 或 MacOS：在 terminal 中运行
  - ./artratex.sh xa：获得全编译后的 PDF 文档
  - ./artratex.sh x：快速编译，不会生成文献引用

全编译指运行 xelatex+bibtex+xelatex+xelatex 以正确生成所有的引用链接，如目录，参考文献及引用等。在写作过程中若无添加新的引用，则可用快速编译，即只运行一遍 LaTeX 编译引擎以减少编译时间。

2.2.3 Tmp 文件夹

运行编译脚本后，编译所生成的文档皆存于 Tmp 文件夹内，包括编译得到的 PDF 文档，其存在是为了保持工作空间的整洁，因为好的心情是很重要的。

2.2.4 Style 文件夹

包含 ucasthesis 文档类的定义文件和配置文件，通过对它们的修改可以实现特定的模版设定。

1. ucasthesis.cls：文档类定义文件，论文的最核心的格式即通过它来定义的。
2. ucasthesis.cfg：文档类配置文件，设定如目录显示为“目 录”而非“目录”。
3. artratex.sty: 常用宏包及文档设定，如参考文献样式、文献引用样式、页眉页脚设定等。这些功能具有开关选项，常只需在 Thesis.tex 中进行启用即可，一般无需修改 artratex.sty 本身。
4. artracom.sty：自定义命令以及添加宏包的推荐放置位置。

2.2.5 Tex 文件夹

文件夹内为论文的所有实体内容，正常情况下，这也是**使用 ucasthesis 撰写学位论文时，主要关注和修改的一个位置，注：所有文件都必须采用 UTF-8 编码，否则编译后将出现乱码文本**，详细分类介绍如下：

- Frontinfo.tex：为论文中英文封面信息。**论文封面会根据英文学位名称如**

Bachelor，Master，Doctor, Postdoctor 自动切换为相应的格式。

- Frontmatter.tex：为论文前言内容如中英文摘要等。
- Mainmatter.tex：索引需要出现的 Chapter。开始写论文时，可以只索引当前章节，以快速编译查看，当论文完成后，再对所有章节进行索引即可。
- Chap_xxx.tex：为论文主体的各章，可根据需要添加和撰写。**添加新章时，可拷贝一个已有的章文件再重命名，以继承文档的 UTF8 编码。**
- Appendix.tex：为附录内容。
- Backmatter.tex：为发表文章信息和致谢部分等。

2.2.6　Img 文件夹

用于放置论文中所需要的图类文件，支持格式有：.jpg, .png, .pdf。其中，ucas_logo.pdf为国科大校徽。不建议为各章节图片建子目录，即使图片众多，若命名规则合理，图片查询亦是十分方便。

2.2.7　Biblio 文件夹

1. ref.bib：参考文献信息库。

2.3　数学公式、图表、参考文献等功能

2.3.1　数学公式

比如 Navier-Stokes 方程（方程 (2.1)）：

$$\begin{cases} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{V}) = 0 \text{ times math test ： } 1,2,3,4,5,1,2,3,4,5 \\ \frac{\partial (\rho \boldsymbol{V})}{\partial t} + \nabla \cdot (\rho \boldsymbol{V}\boldsymbol{V}) = \nabla \cdot \boldsymbol{\sigma} \text{ times text test: 1,2,3,4,5} \\ \frac{\partial (\rho E)}{\partial t} + \nabla \cdot (\rho E \boldsymbol{V}) = \nabla \cdot (k \nabla T) + \nabla \cdot (\boldsymbol{\sigma} \cdot \boldsymbol{V}) \end{cases} \tag{2.1}$$

$$\frac{\partial}{\partial t} \int_{\Omega} u \, \mathrm{d}\Omega + \int_{S} \mathbf{n} \cdot (u \boldsymbol{V}) \, \mathrm{d}S = \dot{\phi} \tag{2.2}$$

$$\mathcal{L}\{f\}(s) = \int_{0^-}^{\infty} f(t) e^{-st} \, \mathrm{d}t, \ \mathscr{L}\{f\}(s) = \int_{0^-}^{\infty} f(t) e^{-st} \, \mathrm{d}t$$

$$\mathcal{F}\big(f(x+x_0)\big) = \mathcal{F}\big(f(x)\big) e^{2\pi i \xi x_0}, \ \mathscr{F}\big(f(x+x_0)\big) = \mathscr{F}\big(f(x)\big) e^{2\pi i \xi x_0}$$

数学公式常用命令请见 WiKibook Mathematics。artracom.sty 中对一些常用数据类型如矢量矩阵等进行了封装，这样的好处是如有一天需要修改矢量的显示形式，只需单独修改 artracom.sty 中的矢量定义即可实现全文档的修改。

2.3.2 数学环境

公理 2.1. 这是一个公理。

定理 2.2. 这是一个定理。

引理 2.3. 这是一个引理。

推论 2.4. 这是一个推论。

断言 2.5. 这是一个断言。

命题 2.6. 这是一个命题。

证明. 这是一个证明。 □

定义 2.1. 这是一个定义。

例 2.1. 这是一个例子。

注. 这是一个注。

2.3.3 表格

请见表 2.1。

表 2.1 这是一个样表。

Table 2.1 This is a sample table.

行号	跨多列的标题							
Row 1	1	2	3	4	5	6	7	8
Row 2	1	2	3	4	5	6	7	8
Row 3	1	2	3	4	5	6	7	8
Row 4	1	2	3	4	5	6	7	8

制图制表的更多范例，请见 ucasthesis 知识小站 和 WiKibook Tables。

2.3.4 图片插入

论文中图片的插入通常分为单图和多图，下面分别加以介绍：

单图插入：假设插入名为c06h06（后缀可以为.jpg、.png、.pdf，下同）的图片，其效果如图 2.1。

图 2.1 Q 判据等值面图，同时测试一下一个很长的标题，比如这真的是一个很长很长很长很长很长很长很长很长的标题。

Figure 2.1 Isocontour of Q criteria, at the same time, this is to test a long title, for instance, this is a really very long very long very long very long very long title.

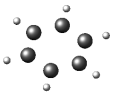

图 2.2 激波圆柱作用。

Figure 2.2 Shock-cylinder interaction.

如果插图的空白区域过大，以图片c06h06为例，自动裁剪如图 2.2。

多图的插入如图 2.3，多图不应在子图中给文本子标题，只要给序号，并在主标题中进行引用说明。

2.3.5　算法

如见算法 1，详细使用方法请参见文档 algorithmicx。

算法 1 Euclid's algorithm

```
1: procedure EUCLID(a, b)                          ▷ The g.c.d. of a and b
2:     r ← a mod b
3:     while r ≠ 0 do                              ▷ We have the answer if r is 0
4:         a ← b
5:         b ← r
6:         r ← a mod b
7:     end while
8:     return b                                    ▷ The gcd is b
9: end procedure
```

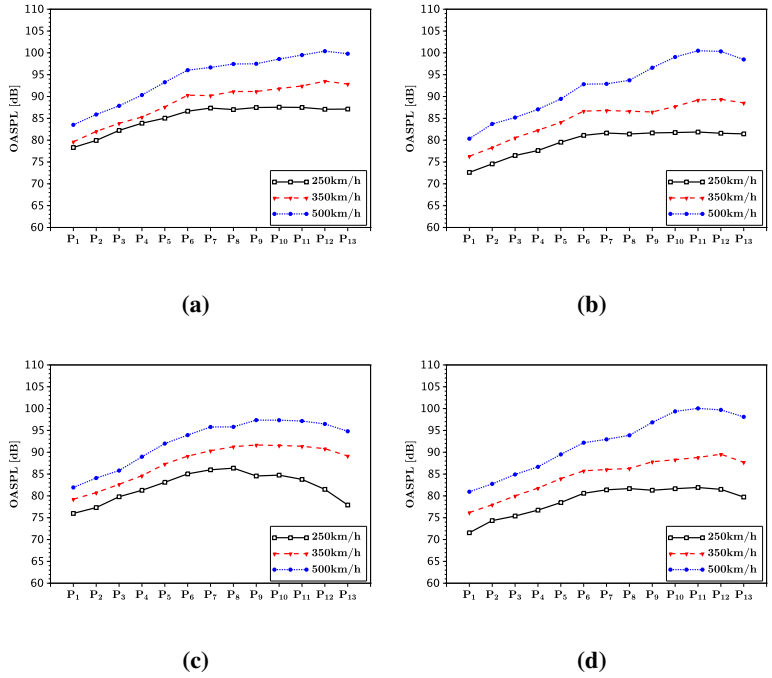

图 2.3 总声压级。(a) 这是子图说明信息，(b) 这是子图说明信息，(c) 这是子图说明信息，(d) 这是子图说明信息。

Figure 2.3 OASPL.(a) This is the explanation of subfig, (b) This is the explanation of subfig, (c) This is the explanation of subfig, (d) This is the explanation of subfig.

2.3.6　参考文献引用

参考文献引用过程以实例进行介绍，假设需要引用名为”Document Preparation System” 的文献，步骤如下：

1）使用 Google Scholar 搜索 Document Preparation System，在目标条目下点击 Cite，展开后选择 Import into BibTeX 打开此文章的 BibTeX 索引信息，将它们 copy 添加到 ref.bib 文件中（此文件位于 Biblio 文件夹下）。

2）索引第一行 @article{lamport1986document,中 lamport1986document 即为此文献的 label (**中文文献也必须使用英文 label**，一般遵照：姓氏拼音 + 年份 + 标题第一字拼音的格式)，想要在论文中索引此文献，有两种索引类型：

文本类型：\citet{lamport1986document}。正如此处所示 Lamport (1986);

括号类型：\citep{lamport1986document}。正如此处所示 (Lamport, 1986)。

多文献索引用英文逗号隔开：

\citep{lamport1986document, chu2004tushu, chen2005zhulu}。正如此处所示 (Lamport, 1986; 初景利, 2004; 陈浩元, 2005)

更多例子如：

Walls 等 (2013) 根据 Betts 等 (2005) 的研究，首次提出...。其中关于... (Walls 等, 2013; Betts 等, 2005)，是当前中国... 得到迅速发展的研究领域 (陈晋镳 等, 1980; Bravo 等, 1990)。引用同一著者在同一年份出版的多篇文献时，在出版年份之后用英文小写字母区别，如：(袁训来 等, 2012a,b,c) 和 袁训来 等 (2012a,b,c)。同一处引用多篇文献时，按出版年份由近及远依次标注。例如 (陈晋镳 等, 1980; Stamerjohanns 等, 2009; 哈里森・沃尔德伦, 2012; 牛志明 等, 2013)。

使用著者-出版年制（authoryear）式参考文献样式时，中文文献必须在 BibTeX 索引信息的 **key** 域（请参考 ref.bib 文件）填写作者姓名的拼音，才能使得文献列表按照拼音排序。参考文献表中的条目（不排序号），先按语种分类排列，语种顺序是：中文、日文、英文、俄文、其他文种。然后，中文按汉语拼音字母顺序排列，日文按第一著者的姓氏笔画排序，西文和俄文按第一著者姓氏首字母顺序排列。如中 (牛志明 等, 2013)、日 (ボハンデ, 1928)、英 (Stamerjohanns 等, 2009)、俄 (Дубровин, 1906)。

如此，即完成了文献的索引，请查看下本文档的参考文献一章，看看是不是

就是这么简单呢？是的，就是这么简单！

不同文献样式和引用样式，如著者-出版年制（authoryear）、顺序编码制（numbers）、上标顺序编码制（super）可在 Thesis.tex 中对 artratex.sty 调用实现，详见 ucasthesis 知识小站之文献样式

参考文献索引的更多知识，请见 WiKibook Bibliography。

2.4 常见使用问题

1. 模板每次发布前，都已在 Windows，Linux，MacOS 系统上测试通过。下载模板后，若编译出现错误，则请见 ucasthesis 知识小站 的 编译指南。

2. 模板文档的编码为 UTF-8 编码。所有文件都必须采用 UTF-8 编码，否则编译后生成的文档将出现乱码文本。若出现文本编辑器无法打开文档或打开文档乱码的问题，请检查编辑器对 UTF-8 编码的支持。如果使用 WinEdt 作为文本编辑器（**不推荐使用**），应在其 Options -> Preferences -> wrapping 选项卡下将两种 Wrapping Modes 中的内容：

TeX;HTML;ANSI;ASCII|DTX...

修改为：TeX;**UTF-8|ACP;**HTML;ANSI;ASCII|DTX...

同时，取消 Options -> Preferences -> Unicode 中的 Enable ANSI Format。

3. 推荐选择 xelatex 或 lualatex 编译引擎编译中文文档。编译脚本的默认设定为 xelatex 编译引擎。你也可以选择不使用脚本编译，如直接使用 LaTeX 文本编辑器编译。注：LaTeX 文本编辑器编译的默认设定为 pdflatex 编译引擎，若选择 xelatex 或 lualatex 编译引擎，请进入下拉菜单选择。为正确生成引用链接和参考文献，需要进行**全编译**。

4. Texmaker 使用简介

(a) 使用 Texmaker “打开 (Open)” Thesis.tex。

(b) 菜单“选项 (Options)” -> “设置当前文档为主文档 (Define as Master Document)”

(c) 菜单“自定义 (User)” -> “自定义命令 (User Commands)” -> “编辑自定义命令 (Edit User Commands)” -> 左侧选择“command 1”，右侧“菜单项 (Menu Item)”填入 Auto Build -> 点击下方“向导 (Wizard)” -> “添加 (Add)”：xelatex + bibtex + xelatex + xelatex + pdf viewer -> 点击“完成 (OK)”

(d) 使用 Auto Build 编译带有未生成引用链接的源文件，可以仅使用 xelatex 编译带有已经正确生成引用链接的源文件。

(e) 编译完成，“查看 (View)” PDF，在 PDF 中“ctrl+click”可链接到相对应的源文件。

5. 模版的设计可能地考虑了适应性。致谢等所有条目都是通过最为通用的

\chapter{item name} and \section*{item name}

来显式实现的 (请观察 Backmatter.tex)，从而可以随意添加，放置，和修改，如同一般章节。对于图表目录名称则可在 ucasthesis.cfg 中进行修改。

6. 设置文档样式: 在 artratex.sty 中搜索关键字定位相应命令，然后修改

(a) 正文行距：启用和设置 \linespread{1.5}，默认 1.5 倍行距。

(b) 参考文献行距：修改 \setlength{\bibsep}{0.0ex}

(c) 目录显示级数：修改 \setcounter{tocdepth}{2}

(d) 文档超链接的颜色及其显示：修改 \hypersetup

7. 文档内字体切换方法：

- 宋体：国科大论文模板 ucasthesis 或 国科大论文模板 ucasthesis
- 粗宋体：**国科大论文模板 ucasthesis** 或 **国科大论文模板 ucasthesis**
- 黑体：**国科大论文模板** ucasthesis 或 **国科大论文模板** ucasthesis
- 粗黑体：**国科大论文模板 ucasthesis** 或 **国科大论文模板 ucasthesis**
- 仿宋：国科大论文模板 ucasthesis 或 国科大论文模板 ucasthesis
- 粗仿宋：国科大论文模板 **ucasthesis** 或 国科大论文模板 **ucasthesis**
- 楷体：国科大论文模板 *ucasthesis* 或 国科大论文模板 *ucasthesis*
- 粗楷体：**国科大论文模板 *ucasthesis*** 或 **国科大论文模板 *ucasthesis***

第 3 章　基于半监督学习的直播电商视频分类网络

3.1　引言

3.1.1　研究背景和动机 Motivation

直播电商内容丰富，监管涉及到不同的部门，面对大量的视频进行监管，单纯靠人工进行分类并分发各个部门去执行，显然是不可能的，因此，对直播电商的监管，首要的是能够对其进行分类，并按照职责进行分发，这是当前面临的最紧迫任务。

3.1.2　相关工作 Relative Work

按照直播电商主体责任，电商平台具有第一位的责任，因此各大电商平台开展了类似的研究，但仅限于本平台内部，淘宝、爱奇艺、美团、京东等都开展了类似的研究。

3.1.3　研究内容和贡献 Our Approach

我们的研究，涉及到辖区内所有平台的直播电商，具有按职责分发、视频识别、营销行为判别、知识产权纠正、在线取证、执法文书自动生成等效果。主要研究内容：

1. 基于半监督学习的直播电商视频分类：近年来，政府在直播电商平台监管实践过程中，按照相关规定，收集建立了直播电商监管数据集，人工分发的标签内容，采取半监督学习的方法对现有直播电商视频进行分类

2. 直播电商营销行为特征库：

3. 基于多模态字典学习的直播电商视频动作识别：采取字典学习的方法对直播电商营销活动进行动作识别

4. 营销行为判别

(a) 基于声音 + 视频：双击运行 artratex.bat 脚本。

(b) 基于视频 + 文本: terminal -> chmod +x ./artratex.sh -> ./artratex.sh xa

(c) 基于多模态：都可使用 LaTeX 编辑器打开 Thesis.tex 文件并选择 xelatex 编译引擎进行编译。

5. 异常处理：知识产权侵权检测

6. 在线取证

7. 执法文书自动生成（视频结构化描述）

3.2 论文主要结构

绪论，简要概括了课题研究背景、目的和论文研究内容，设计和规划了论文的整体框架。

3.2.1 第 2 节

对过去 5 年来的历史文献进行回顾，从多个视角描述了直播电商监管的发展历程，总结了技术上取得的进展和存在的困难和问题，提出本文开展研究的目标和内容。

3.2.2 第 3 节

提出了基于半监督学习的直播电商视频分类。

第 4 章　Chapter Two

参阅论文 Long-Term Feature Banks for Detailed Video Understanding ucasthesis 参阅论文

摘要

为了更好地理解直播电商视频中营销事件，需要将视频正在发生的行为与过去联系起来，并将营销事件置于上下文中。在本节中，为了使现有的视频模型也能做到这一点，借鉴长期特征库的思想，提出一个直播电商营销行为长期库，利用过去直播电商视频监管过程中提取的支持信息，以增强当前最先进的视频模型，有效弥补这些模型只能查看 5 秒以内的短片。实验表明，使用长期特征库增强 3D 卷积网络可以在四个具有挑战性的视频数据集上产生最先进的结果：AVA、EPIC-Kitchens 和 Charades，并在现有的直播电商监管数据集上达到 81.7%。

4.1　引言

4.1.1　Motivation

直播电商监管的关键是如何识别视频中的营销事件，由于营销事件是个典型的复杂事件，因此要想正确、快速识别这些事件，这需要将视频现在发生的事情与过去发生的事情联系起来。如果没有利用过去来理解现在的能力，作为监管人员，是无法理解我们正在监管的内容。

4.1.2　Relative Work

长期以来，计算机视觉研究中使用 ImageNet 预训练网络从孤立的帧中提取特征，然后将这些特征用作训练池或循环网络的输入，这些相同的特征既反映当前的背景，也表达了长期信息。[Long-term Feature Bank] 通过使用预先计算的视觉特征来利用长期时间信息这一理念 [25, 31, 45, 57]，提出了一个长期特征库的想法，该特征库存储整个视频的丰富的时间索引表示，将过去和可用的未来场景、对象和动作的信息进行编码并存储形成长期特征库。这些信息支持上下文内容，允许视频模型（例如 3D 卷积网络）更好地推断当前正在发生的事情。现有的实验证明，该长期特征库能够改进最先进的视频模型，克服了大多数预测仅基

于来自短视频剪辑信息的不足，有效解决 3D 卷积端到端网络必须密集采样才能有效工作和视频输入片段较短的问题。同时将当前信息与长期信息解耦，将长期特征库视为增强标准视频模型的辅助组件，例如最先进的 3D CNN。这种设计使长期特征库能够存储灵活的支持信息，例如与 3D CNN 计算的不同的对象检测特征。

4.1.3 Our Approach

本节借鉴长期特征库思想，提出了一个基于直播电商监管营销行为特征库。在实践应用中，这个营销行为特征库可与 3D CNN 简单集成起来。实验证明了多种机制是可行的，包括一种注意力机制，它将关于当前的信息（来自 3D CNN）与存储在长期特征库中的远程信息相关联。在对象级以及帧或视频级预测的数据集上结果表明，这个特征库可在具有不同输出要求的不同任务中得以应用。最后，大量的实验说明，使用营销行为特征库增强 3D CNN 在直播电商监管视频视频数据集上产生了最先进的结果。消融研究也表明，这些任务的改进源于长期信息的整合。

本节结构如下：

1. 第 2 节介绍了相关的工作。
2. 第 3 节介绍了营销行为特征模型
3. 第 4 节在直播电商监管数据集上的试验

(a) 实现细节

(b) 消融实验

4. 第 5 节介绍了与 SOTA 比较
5. 最后小结。

4.2 相关工作

4.2.1 深度神经网络 Deep networks

深度神经网络是视频理解的主要方法 [5、21、33、39、46-48、50、51、56]。既有应用广泛的双流网络 [21,39,50] 和 3D 卷积网络 [5,33,46-48,51,56]。在本节采用 3D CNN，但营销行为特征库也可以与其它的视频模型很方便地集成使用。

4.2.2　时间关系模型 Temporal and relationship models

时间关系模型包括建模视频帧演化的 RNN[7,24,27,44,57] 和建模有序帧特征 [58] 的多层感知器（MLP）。为了建模更细粒度的交互，越来越多的工作线利用预先计算的对象提取 [52] 或检测 [4,30]，并在一个短片段内对它们的共现 [30,43,52]、时间顺序 [4] 或空间排列 [52] 进行建模。

4.2.3　长期视频理解 Long-term video Understanding

目前,CNN 对长期视频理解研究较少，已知部分原因是由于硬件的限制，如 GPU 内存。现有的研究表明，使用预先计算的特征是克服这些限制的一个有效办法，端到端训练 [25,31,45,57] 还需要关键的突破。但由于这些方法并不能优化目标任务的特征，因此结果仅可能是次优的。另一种方法是使用 aggressive subsampling[50,58] 或大大步 [8]。TSN[50] 采样每段视频 3-7 帧。ST-ResNet[8] 使用的时间步幅为 15。本文采取端到端学习 strong 短期特征、密集采样和灵活解耦，以及灵活的长期建模，据现有的实验表明，本文采取的方法是最优的。

4.2.4　时空行为定位 Temporal action localization

时空动作定位是当前计算机视觉领域的一个研究热点 [12,14,17,32,40,53]。最近的研究进展是扩展了对象检测框架 [10,34]，首先在短剪辑/框架中提出管/盒子，然后将管/盒子分类为动作类 [14,17,20,32,36,36]。检测到的管/盒可以选择连接形成完整的作用管 [12,17,20,32,36,40]。这些方法可以在每一帧或剪辑中独立地找到动作，而不需要利用长期的上下文。而 [Long-term feature bank] 使用的方法与上相反。

4.2.5　信息库 Information bank

已知的对象库 [26]、检测库 [1] 和存储器网络 [42] 等信息特征库表示，已在图像级广泛使用，在视频索引和检索、文本语料库中的信息建模等得到积极应用。本节从上述方法中获得灵感，为具体的视频理解任务探索简单而又方便的方法。

4.3 直播电商营销行为特征库模型

现有的计算机视觉模型表明，要想对长而复杂的视频做出准确的预测，必须将当前发生的事情与时间上遥远的事件联系起来。具体对直播电商监管来说，营销活动是一个复杂事件，要想对长而复杂的营销活动做出准确的识别，必须考虑直播电商营销行为长期特征。针对这一想法，本文提出了一个带有长期特性库的直播电商营销行为特征库模型，以实现长期视频理解进程中的交互。

4.3.1 方法概述

本节描述了如何用于时空行动定位任务的方法，目标之一是检测视频中的网络主播，并对他们的行为进行分类。以前大多数最先进的方法 [9,14,43] 是把一个“主干”的 3DCNN(例如，C3D[46]，I3D[5]) 与一个基于区域的人探测器 (例如，Fast/Faster R-CNN[10,34]) 结合起来。为了处理一个视频，它被分成 2-5 秒的短片，通过 3DCNN 独立转发计算一个特征图，然后与区域建议和感兴趣区域 (RoI) 池一起计算每个候选参与者 [9,14] 的 RoI 特征。这种方法仅能捕获短期信息，如图 3a 所示。

本节使用的方法，在 long-term feature bank 上加以扩展应用：（1）长期特征库。直观地作为整个视频中发生的事情的“记忆”，将计算定期采样时间步长的 RoI 特征；（2）是一个特征库操作符 (FBO)，它计算短期行为特征（描述参与者现在正在做什么）和长期特征之间的交互。（3）交互可以通过注意机制计算，如非局部块 [51]，或通过特征池和连接。我们的模型总结在图 3b 中。下面，将详细介绍这个办法。

4.3.2 直播电商营销行为特征库

直播电商营销行为特征库，L，目的是提供相关的上下文信息，以实现在当前的时间步长中进行行为识别。当用于时空动作定位的任务，如具体营销行为时，需要在整个视频上运行一个网络主播行为的探测器，为每一帧生成一组检测。同时，运行一个标准的基于剪辑的 3DCNN，如 C3D[46] 或 I3D[5]，在视频上以有规则的间隔间隔，比如每一秒一次。然后我们使用 RoI 池来提取使用 3D CNN 处理的每个时间步对所有人的检测特征。形式上，L=[L0，L1，...，LT−1] 是视频时间步长 0，...，T−1 的时间索引特征列表，其中 Lt∈RNt× 是时间 t× 维 RoI 特征的矩阵。直观地说，L 提供了关于整个视频中所有人员在何时以及做什

么的信息，并且可以通过探测器和 3D CNN 通过视频一次来有效地计算出来。

4.3.3　直播电商营销行为特征操作

本节提出的模型，通过营销行为特征操作符 FBO(St，˜Lt) 引用了来自长期特征 L 的信息。特征库操作符接受输入圣 ˜Lt，圣是短期投资回报率汇集特性和 ˜Lt[Lt−w，...，Lt+w]，L 的一个子集集中在当前剪辑在 t 窗口大小 2w+1，堆积成矩阵 ˜Lt∈RN×d，N=t+wtwtwNt。我们将窗口大小 2w+1 作为一个我们在实验中交叉验证的超参数。然后，输出通过通道与 St 连接，并用作线性分类器的输入。直观地说，特征银行操作符通过将其与长期特征联系起来，计算出合并的短期特征 St 的更新版本。FBO 的实现是灵活的。注意机制的变体是一个明显的选择，我们将在实验中考虑多个实例。

4.3.4　实现细节

4.3.4.1　主干

我们使用了一个标准的 3D CNN 架构作为主干。该模型是一个 ResNet-50[16]，在 ImageNet[35] 上进行预训练，并使用 I3D 技术 [5]“扩展”成一个具有 3D 卷积（超过空间和时间）的网络。网络结构被修改为包括非局部操作 [51]。在将网络从二维扩展到三维后，我们在动力学-400 数据集 [5] 上进行视频分类的预训练。该模型在动力学-400[5] 验证集上达到了 74.9%（91.6%）的前 1 名（前 5 名）的精度。最后，我们在 [52] 之后删除了 conv1 和 pool1 的时间步进，并删除了特定于动力学的分类层，以生成主干模型。确切的模型规格详见补充材料。所得到的网络接受形状为 32×H×W×3 的输入，代表 32 个具有 H×W 空间大小的 RGB 帧，并输出形状为 16×H/16×W/16×2048 的特征。相同的体系结构用于计算短期特征 S 和长期特征 l。除非另有说明，否则参数在这两个模型之间不共享。

4.3.4.2　RoI 池

本文首先在时间轴上对视频主干特征进行平均池。然后，我们使用空间输出为 7×7 的 RoIAlign[15]，然后使用空间最大池化，为 RoI 生成一个单一的 2048 维特征向量。这相当于使用一个时间直管 [14]。

4.3.4.3 使用

直播电商营销行为特征操作安装。可通过多种方式来使用直播电商营销行为特征库的操作。-LFB NL：我们的默认特征库算子 FBONL(St,L t) 是注意力算子。直观地说，我们使用 St 来关注 L t 中的特征，并通过快捷连接将关注的信息添加回 St。我们使用一个简单的实现，其中 FBONL(St,L t) 是最多三个非局部 (NL) 块的堆栈 [51]。我们用局部特征 St 和长期特征窗口 L t 之间的注意力替换标准非局部块 [51] 的自注意力。此外，我们的设计使用层归一化（LN）[3] 和 dropout [41] 来改进正则化。由于我们的目标任务包含相对较少的训练视频，在训练时容易表现出过度拟合，因此使用正则化修改对于实际应用非常重要。修改后的堆栈，其中 θ1,2,... 是可学习的参数。类似于王等人 [52]，我们使用线性层将 FBONL 输入维度降低到 512，并以 0.2 的速率应用 dropout [41]。因此最终线性分类器的输入是 2048 (St) 512 (FBONL 输出) = 2560 维。

4.3.4.4 训练

由于通过长期特征库进行反向传播的计算和内存复杂性，对整个模型的端到端联合训练 (图 3b) 是不可行的。相反，我们将用于计算 L 的 3DCNN 和检测器视为固定组件，它们离线训练，但仍然在目标数据集上，随后不会更新。我们实验了交替优化方法来更新这些模型，类似于目标传播 [23]，但发现它们并没有改善结果。稍后将给出特定于数据集的培训细节。

4.3.4.5 短期操作基准

为了验证合并长期信息的好处，此外，我们还研究了该模型的“退化”版本。相反，它使用了一个与 FBONL 相同的短期操作符，但只引用了来自剪辑中的信息：STO(St)：=FBONL(St，St)。STO 在概念上类似于 [52]，并允许反向传播。我们观察到与 STO 的大量过拟合，因此应用了额外的正则化技术。详见补充资料。

4.4 实验

本文使用 AVA 数据集 [14] 进行广泛的消融研究。AVA 由 235 个训练视频和 64 个验证视频组成；每个视频都是一个 15 分钟的视频片段。帧以 1FPS 稀疏地标记。标签是：框中每个人周围的一个边界框，以及一个多标签注释，指定框中的人在标记帧的 ±0.5 秒内从事哪些操作。操作标签空间被定义为由数据集作

者定义的 80 个“原子”操作。AVA 中的任务是时空动作定位：出现在测试视频中的每个人都必须在测试视频的每一帧中被检测到，并且必须正确预测被检测到的人的多标签动作。算法的质量是由一个平均平均精度 (mAP) 度量来判断的，该度量需要至少 50% 的联合 (IoU) 重叠的交集，以使检测与地面真相相匹配，同时预测正确的操作。

4.4.1　实现细节

接下来，我们将描述了用于 AVA 的对象检测器、输入采样以及训练和推理细节。

4.4.1.1　网络主播检测

本文使用 Faster R-CNN[34] 和 ResNeXt-101-FPN[28,55] 作为主干来进行网络主播的检测。该模型在 ImageneNet[35] 和 COCO 关键点 [29] 上进行预训练，然后在 AVA 边界框上进行微调；训练细节见补充材料。最终的模型在 AVA 验证集上获得 93.9AP@50。

4.4.1.2　时间采样

短期和长期特征都是由 3D CNN 提取的，使用 32 帧采样，时间步幅跨越 63 帧 (30FPS 视频中 ~2 秒)。长期特征在整个视频中以每秒一个剪辑的速度计算，并在 AVA 上对 3DCNN 模型 (图 3a) 进行了微调。

4.4.1.3　训练

实验中使用同步 SGD 来训练现有的模型，在 8 个 GPU 上使用 16 个剪辑 (即每个 GPU 2 个剪辑)，与批标准化 [18] 层冻结。对所有模型进行了 140k 次迭代的训练，学习率为 0.04，在迭代 100k 和 120k 时降低了 10 倍。权重衰减为 10^{-6} 0.9 $224 \in [256, 320] \times 224$ 0.9 Io

4.4.1.4　推理

在测试时，我们使用 ≥0.85 的检测。所有模型都将短边重新缩放到 256 像素，并使用单中心裁剪 256×256。对于训练和推理，如果一个框跨越裁剪边界，我们将裁剪剪辑中的区域池。在罕见的情况下，一个盒子掉出裁剪区域，RoIAlign[15] 在边界处汇集该特征。

图 6 显示了使用不同窗口大小的 LFB 的相对增益。我们看到不同的数据集表现出不同的特征。视频数据集 AVA 得益于持续 2+ 分钟的非常长的上下文。实现对直播电商的监管，正确识别直播电商营销行为是必要的，而且识别 15 至 60 秒的上下文内容，甚至更长时间的营销活动内容尤其重要。Charades 视频要短得多（~ 为 30 秒），但将时间支持扩展到 10+ 秒仍然是有益的。可以预见，未来更具挑战性的数据集可能会受益更多。

综上所述，为了更好地识别直播电商视频中的营销行为，我们提出了一个用于为视频模型提供长期支持性信息的直播电商营销行为长期特征库。我们展示了，通过 LFB 使视频模型能够访问长期信息，可以带来巨大的性能提高，并在 AVA、EPIC-Kicchens 和 Charades 等具有挑战性的数据集上产生最先进的结果。

第 5 章　MSL-VID-2019 数据集: 基于直播电商监管业务视频数据集构建

摘要直播电商进入爆发期，给监管工作带来颠覆性压力，如何加强直播电商监管成为全国各级监管部门的紧迫性课题。直播电商主要以视频的形式为载体，但具有区别于传统视频的特性。虽然目前对视频研究的数据集不少，但鲜有直播电商的视频数据集来支持。本文介绍了我们在研究直播电商监管业务时构建的视频数据集：MSL-VID-2019 数据集在当前版本中，MSL-VID-2019 提供了 203 个活动类的样本，平均每个类 137 个未修剪的视频，每个视频 1.41 个活动实例，总共 849 个视频小时，旨在研究如何解决电商直播视频实际场景下的识别问题。

5.1　引言

数字经济扑面而来，直播电商进入爆发期，作为新业态在成为当前经济发展的亮点的同时，给政府监管部门带来了新的课题，也给监管工作带来颠覆性压力。如何加强直播电商监管成为全国各级监管部门的紧迫性课题。电商是直播视频流量变现的理想场景，直播电商的载体主要是直播视频，因此加强对直播电商的直播视频进行研究成为当前监管的关键。

随着直播电商用户和活动增长，直播电商视频数据库的数量和规模都在以令人难以置信的速度增长。尽管视频数据的爆炸式增长，但自动识别和理解电商网络交易行为活动的能力仍然相当有限。对现有的监管人员来说，阻碍当前技术性能研究的一个重要限制，是存在的视频数据集问题，即缺乏可以解决电商直播视频实际场景下的视频数据集和研究标准。现有的数据库往往是具体的，并专注于有限类型的活动。

本文构建了一个数据集 MSL-VID-2019。该数据集围绕一个直播电商营销语义本体构建，语义本体根据电商平台直播营销行为来指导组织活动。它具有至少三个深度层次的结构，是一个在丰富的语义分类下组织的直播电商网络交易行为识别数据库。

本文组织结构如下：首先，我们回顾和总结了现有的直播电商网络营销的标准。然后，详细介绍了我们收集的数据集和标注的细节，并提供了对 MSL-VID-2019

的属性的总括。

5.2 相关工作

在过去的十几年里，各国学者提出了最具影响力的行动数据集。

5.2.1 最初的视频数据集

如 Hollywood dataset[20] 包含了取自好莱坞电影的视频，由 12 个动作类别由专业演员执行，这比早期的简单动作数据集 [33,9] 产生更自然的场景。类似地，UCF Sports 数据集 [30] 和 Olympic Sports 数据集 [24] 通过关注高度清晰的体育活动，增加了动作的复杂性。但是，类别的数量较少，使得活动的范围较窄。复杂性的另一个维度是由关注可组合 [21] 和并发 [41] 活动的数据集来解决的，但这些活动受到场景和环境假设的约束。

5.2.2 UCF101[17]-Thumos 14[35] 和 HMDB51[19] 数据集

这些数据集由 YouTube 视频编译，有 50 多个动作类别。所得到的视频样本很短，而且只传达了简单的短期行动或事件。这些视频是通过手工和昂贵的过程收集的，如果要扩展数据集的大小，就很难进行缩放。在语义组织方面，HMDB51 将活动分为 5 种主要类型：一般面部、与物体操作的面部、一般身体运动、与物体互动的身体运动和与人体互动的身体运动。另一方面，UCF101 将类别分为 5 类：人物互动、身体运动、演奏乐器、运动。不足的是，这些是简单的分类，只有两个级别的解决，没有提供详细的活动事件。

5.2.3 MPII 人体姿态数据集 [2]

专注于人体姿态估计，最近被应用于动作识别 [29]。它提供了描绘人类行为的短片段（41 帧或更长的时间）。不足的是，每个类别的视频样本的分布是不均匀的，且偏向于某些行动类别。

目前，可用的最大的视频数据集是 Sports-1M 数据集 [16]，大约有 500 个体育相关类别，由自动标记算法注释。尽管该数据集规模庞大，但该数据集使用了一些有限的活动分类法，因为它只关注体育行为。此外，自动收集过程引入了一个未公开量的标签噪声。

与我们的工作相关的还有努力为静态图像中的对象识别构建大尺度的基准。诸如 ImageNet[5]、SUN[42] 和 TinyImages[36] 等图像基准测试已经在计算机视

觉算法的相关任务中取得了重大进展。大规模视觉识别挑战 (ILSVRC)[32] 就是一个例子，AlexNet 架构 [18] 由于在挑战中的出色表现而广受到欢迎。

5.2.4　ActivityNet

涵盖了与人类如何在日常生活中花费时间最相关的活动；每个视频的数量和长度（而不是短片段）、活动分类的多样性和类的数量的定性跳跃；与全自动注释算法相比，人的循环注释过程可以提供更高的标签精度；以及一个低成本连续数据集扩展的框架。

5.2.5　Kinetics

人体动作视频数据集。该数据集包含 400 个人体动作类，每个动作至少有 400 个视频剪辑。每个片段持续约 10 秒，取自 YouTube，包含了大量的行为特征，但行为较为简单。

综上所述，大多数视频数据集关注的模态特征有限，对人类行为的涉及简单，数据规模较短，没有可以针对解决多模态人物行为识别问题。

本文介绍的 MSL-VID-2019 数据集，是用于多模态电商直播视频识别的大规模数据集。与表 1.1 中列举了一些当期流行的数据集进行对比可知，这是一个当前最大的多模态电商直播视频监管数据集。其总时长超过 5230 小时，共计 70 多万个视频。该数据集是从广泛类型的实际电商平台在线视频中提取的视频片段。然后，监管人员用人工注释标记了少部分剪辑，并采用自动算法以加快收集和标记过程。首先从庞大的长视频数据集中提取视频剪辑。然后，通过自动算法过滤掉没有人或多人的剪辑。最后，按照身份对候选片段进行分组，然后将其放入手动注释中。

5.3　MSL-VID-2019 电商直播视频介绍

MSL-VID-2019 数据集的目的是提供一个描述直播电商网络交易行为视频的语义组织。在本节中，我们将介绍作为 MSL-VID-2019 的主干的定义电商直播营销的行为词典和层次结构。然后介绍了电商直播视频的收集和标注。

5.3.1　直播电商营销的行为字典

MSL-VID-2019 数据集是用于多模态直播电商营销行为识别的大规模数据集。电子商务是基于网络交易语义的一种行为，我们基于分类法的基础上构建

MSL-VID-2019 数据集。我们依据《电子商务参与方分类与编码》(中国物品编码中心, 北京交通大学 et al. 2016) 和《电子商务信用　网络交易信用主体分类》(中国标准化研究院, 深圳市众信电子商务交易保障促进中心 et al. 2015) 国家标准，对层次结构提供的 4000 多个活动示例中手动选择了 203 个活动类别的子集。活动类属于 2 个不同的顶级类别，9 个二级类别，我们还准备将数据集用分类法刻画四个粒度级别，它构成了一个语义组织主干，在模型训练期间将对利用层次结构的算法中有很大的好处。

5.3.2 直播电商视频收集和预处理

根据国家市场监管总局 37 号令《网络交易监督管理办法》（以下简称《办法》），直播服务提供者需将网络交易活动的直播视频自直播结束之日起至少保存三年，网络交易新业态的经营者需以显著方式展示商品或者服务及其实际经营主体、售后服务等信息，充分保障消费者的知情权，县级以上地方市场监督管理部门应当在日常管理和执法活动中监管部门对直播服务提供者开展监督检查、案件调查、事故处置、缺陷消费品召回、消费争议处理，并实施信用监管。视频集来自全国市场监管部门在日常监管、抽检监测、定向监测和双随机抽查过程中对监管的主要 19 个直播电商平台保存内容的采样。(国家市场监督管理总局网络商品交易监督管理司 2021.3) 这些电商直播视频剪辑是依据《直播营销管理规定》，按照电商直播电商平台监管规定和要求，由监管人员在日常监督检查中获得的。从全国各地区现有电商直播视频平台中大量的在线视频中提取的。

为了使数据集更接近真实情况，该数据集是从广泛类型的真实在线视频中提取的视频片段。然后，监管人员用人工注释标记了部分剪辑监管分类标注，并采用自动算法以加快收集和标记过程，该操作的流程图如图 1.5 所示。首先从庞大的长视频数据集中提取视频剪辑。然后，通过自动算法过滤掉没有人或多人的剪辑。之后，按照身份对候选片段进行分组，然后将其放入手动注释中。

5.3.2.1 提取视频片段

原始视频是市场监管部门从电商直播视频数据库提取得，涵盖电商直播营销的等各个方面。根据连续帧之间的差异，每个原始视频将被分割为多个镜头。短于 3 秒钟的视频剪辑会从数据集中排除，因为它们通常缺少多模态信息。本文取 CNN 作为电商直播视频输入的一个原始框架，并将其传递给卷积层来理解直

播视频状态。由于直播电商视频原始帧将有 210x160 像素和 128 色的调色板，如果采取直接输入原始像素的方法，会耗费的大量的无法承担的计算和内存。在试验环境中，我们采取像素降维到 84x84，同时将 RGB 值转换为灰度值。采取这样预处理后，我们把电商直播视频素材作为卷积层的输入。我们使用两个卷积层，然后是一个以 ReLU 作为激活函数的全连接层。在这里，我们不使用池化层。当我们执行对象检测或分类等任务时，池化层的作用至关重要，其中我们不考虑对象在图像中的位置，只要确定所需的对象在图像中即可。

5.3.2.2　通过头部检测自动过滤和主播身份获取候选剪辑

作为识别人员的标准，每个视频剪辑都必须包含一个且只有一个主要人物。为了找出主要人物，使用 YOLO V2 算法可检测到头部区域。有效帧定义为这一帧画面中仅检测到一个头像区域的帧，或者最大头像区域的面积比其他头像区域的面积大三倍的帧。有效片段定义为有效帧超过 30% 的比例的片段。不包含有效帧的剪辑被称为无效的剪辑，将被删除。但是，由于头部检测器无法检测到所有头部，因此在此阶段可以保留一些片段。一些噪声片段将在手动过滤步骤中被丢弃。对包含有多个人的视频剪辑片段会使数据集更接近实际应用程序，但这也大大识别和标注的难度，在实际应用加以剪切。该视频剪辑包含来自人脸识别数据库的 3253 个主播身份标识的标签。在此阶段，每个剪辑都将通过人脸识别标记一个初始身份，初始身份均是从网络监管的主播数据库中选取。面部由 SSH 模型 [13] 检测，然后再由 Arc Face[3] 模型识别，并对营销活动标注行为标签。

5.3.2.3　噪声检测和过滤

在导入的视频中反复用双流动作分类器 [19] 来识别这些噪声。将检测到的混淆来合并、分割或完全删除。在收集了所有数据、删除重复数据后，最后的手动剪辑过滤阶段，使用双流模型的类进行打分，分数排列最低的即是噪声。同时可利用排列找到毗邻的类余的重复视频，并过滤掉那些视频。

5.3.2.4　手动标注监管目标分类

约 30% 剪辑片段都通过手动注释过程，针对监管职责的分工要求，对监管类别进行标注。后续将针对监管类目进行标注。通过数据清理后，系统还会随机选择了 10％的数据集进行质量测试，数据标签错误率保持在 0.2％以内。

5.3.2.5 数据统计

该 MSL-VID-2019 数据集包含了 65 万个视频片段，整个数据集由｛图像，主播，商品名称，监管分类｝四元组组成，主播姓名由短语组成，商品名称和监管分类来自相应字典编码，图像是 base64 编码的 JPEG 图像缩略图，监管分类由人工标注，大约有 1.2 万左右。整个数据集被划分为三个部分，数据集的 40% 用作训练集，数据集的 30 用做验证集，剩下的 30 视频剪辑用于测试集。训练集有 260568 个视频，验证集有 195426 个视频，测试集有 195426 个视频，总计 470 个不同营销场景。通过把该数据集和其他数据集进行对比，发现该数据集视频数目多，营销行为种类多，识别难度大。

表 5.1 不同数据集性能比较

Table 5.1 Compairison of video datasets.

Dataset	Year	Actions Clips	Total	Videos	
HMDB-51 [15]	2011	51	min102	6266	3312
UCF-101	2012	101	min101	13320	2500
ActivityNet-200	2015	200	avg101	28108	
Kinetics	2017	400	min400	306245	306245
MSL-VID-2019	2019	470	min619	572391	651420

5.4 试验

5.4.1 评价指标

为了准确评价实验中所用方法提取倒的特征的质量，需要将它们输入到分类器中进行判别。分类结果的评价标准可以用来评价特征的性能。其评价标准用 Mean Average Precision(m AP)。具体表达如下公式所示.

5.4.2 行为检测

行为检测是应用有关的算法从视频开始帧到结束帧，计算出在视频中出现的每个活动的持续时间。为了评估不同的分类模型，我们利用 MSL-VID-2019 的部分标注实施评估，从而形成电商直播营销活动视频数据集。数据集：我们将现有视频数据集与 ActivityNet 数据集和 Kinetics 数据集进行了对比。在跨模态视频数据集中，我们取得较好的分数。请见表 5.1。