ICS 25.160.01
CCS J 33

CWA

团　　体　　标　　准

T/CWAN 0172—2025

焊接大模型多源数据规范

Multi-Source data specification of welding large-scale model

2025-11-17 发布　　2025-12-01 实施

中国焊接协会　发布

目　　录

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国焊接协会提出并归口。

本文件起草单位：北京博清科技有限公司、中国科学院合肥物质科学研究院、中国兵器工业集团航空弹药研究院有限公司、天津市特种设备监督检验技术研究院、清华大学、上海中巽科技股份有限公司、广西柳工机械股份有限公司、北部湾大学、北京理工大学（珠海）、哈尔滨华德学院、福建省特种设备检验研究院、哈尔滨职业技术大学、哈尔滨中焊协学技术服务有限公司、坤智大数据科技（哈尔滨）有限公司。

本文件主要起草人：冯消冰、韩滕跃、李杰庆、韩冬、贺柏达、王铭秋、刘爱平、程启超、徐玉平、彭吴擎亮、侯国清、孙明辉、黎欣、范东辉、李长威、于兴华、王万景、郝亮、马青军、王贵锦、王滨滨、苏楠、崔元彪、方乃文。

焊接大模型多源数据规范

1 范围

本文件规定了焊接大模型多源数据，涵盖数据的分类、数据采集、数据处理、数据标注及数据隐私与安全等数据的全生命周期过程。

本文件适用于开发、部署和使用焊接大模型的教育单位、科研机构、企业、技术开发者及相关组织等。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 3375 焊接术语

GB/T 36344—2018 信息技术 数据质量评价指标

GB/T 37973—2019 信息安全技术 大数据安全管理指南

GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型

GB/T 42127—2022 智能制造 工业数据 采集规范

GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

3 术语和定义

GB/T 3375 中界定的及下列术语和定义适用于本文件。

3.1

坡口图像 laser image of groove

通过工业相机捕捉到激光发射器发射线的激光条纹投射在目标焊缝上并在焊缝表面形成漫反射的坡口图像。

3.2

熔池图像 weld pool image

通过视觉传感器（如高速摄像机、红外热像仪等）实时采集的熔池动态形貌的数字图像。这些图像包含熔池的几何特征（如长度、宽度）、温度分布和流动行为等信息。

3.3

坡口视频 laser video of groove

焊接设备沿着焊缝运动过程中，搭载到焊接设备上的工业相机以一定的频率采集坡口图像而形成的视频。

3.4

熔池视频 melting pool video

焊接设备沿着焊缝运动过程中，搭载到焊接设备上的工业相机以一定的频率采集熔池图像而形成的

视频。

3.5

电弧声音 arc sound

在焊接过程中，由电弧放电、熔滴过渡、等离子体振荡和熔池动态行为等物理现象产生的可听或高频声波信号。这些声音信号包含了焊接工艺状态、质量缺陷和能量传递的关键信息，可用于实时监测、工艺优化和智能控制。

3.6

元数据 metadata

关于数据或数据元素的数据（可能包括其数据描述），以及关于数据拥有权、存取路径、访问权和数据易变性的数据。

[来源: GB/T 36344—2018，2.2]

4 数据分类

4.1 文本数据

可用于焊接大模型预训练、微调的文本数据包含但不限于以下内容：

a) 焊接领域相关书籍、标准、规范、文献等数据；

b) 焊接作业过程中产生的焊接工艺参数数据；

c) 焊接工艺规程数据；

d) 焊接接头检测报告数据。

4.2 图像数据

可用于焊接大模型预训练、微调的图像数据包含但不限于以下内容：

a) 坡口图像数据；

b) 熔池图像数据；

c) 坡口示意图像；

d) 焊接接头成型图；

e) 焊接接头质量检测图。

4.3 视频数据

可用于焊接大模型预训练、微调的视频数据包含但不限于以下内容：

a) 坡口视频数据；

b) 熔池视频数据。

4.4 音频数据

可用于焊接大模型预训练、微调的音频数据包含但不限于以下内容：

a) 焊接过程中产生的声音数据。

4.5 其他数据

a) 具有特定格式的文件，例如 3D 模型、点云 PLY 等。

5 数据采集

数据采集宜总体上遵循 GB/T 42127—2022 智能制造 工业数据 采集规范中的要求。

5.1 数据采集要求

5.1.1 文本数据采集要求

a) 焊接领域相关书籍、标准、规范、文献等数据采集要求

1) 应严格遵守版权法规，规范引用和标注；

2) 标注来源信息： 对每一条采集的数据，必须清晰、完整、规范地记录其来源信息，如下：

书籍：作者、书名全称、出版版次、出版社、出版年份、国际标准书号（ISBN）、页码等。

标准：中国标准分类号、国际标准分类号、发布日期、实施日期、主管部门、归口部门、发布单位等。

文献：第一作者、标题、期刊/会议名称、卷号/期号/会议信息、页码、出版年、数字对象唯一标识符（DOI 号）等。

b) 焊接作业过程中产生的焊接工艺参数数据采集要求

1）宜记录产生该数据的项目信息（项目名称或项目编号）、产品信息（产品名称或产品编号）、焊缝编号、采集时间、母材信息、焊接方法、焊接材料信息、焊接电源型号、外部环境温度、湿度、风速等数据。

2）焊接电流：数据类型为 int64，单位为安培(A)，按时间要求采集频率≥20 Hz 或者按焊接里程要求采集频率≤1 mm。

3) 焊接电压：数据类型为 float64，单位为伏特(V), 按时间要求采集频率≥20 Hz 或者按焊接里程要求采集频率≤1 mm。

4) 焊接速度：焊接设备沿焊接方向移动的速度，数据类型为 int64，单位为毫米每分钟(mm/min), 按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

5) 送丝速度：送丝速度为单位时间内焊丝向焊接熔池送进的长度，数据类型为 int64，单位为米每分钟(m/min)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

6) 摆幅：焊接摆幅为焊接过程中焊枪沿垂直于焊接方向的横向移动范围，数据类型为 float64，单位为毫米(mm)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

7) 左摆速：左摆速为焊枪或焊丝在摆动焊接过程中向左侧移动时的速度，数据类型为 int64，单位为毫米每秒(mm/s)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

8) 右摆速：右摆速为焊枪或焊丝在摆动焊接过程中向右侧移动时的速度，数据类型为 int64，单位为毫米每秒(mm/s)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

9) 左停留时间：时间为焊接过程中焊枪在焊缝的左侧停留的时间，数据类型为 float64，单位为秒(s)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

10) 右停留时间：时间为焊接过程中焊枪在焊缝的右侧停留的时间，数据类型为 float64，单位为秒(s)，按时间要求采集频率≥20Hz 或者按焊接里程要求采集频率≤1mm。

c) 焊接工艺规程数据

1）焊接工艺规程列表：包含的要素为工艺规程、适用项目、焊接工艺、支持的工艺评定编号、钢级、管径、壁厚等。

2）母材信息：包含的要素为母材材质、母材厚度、母材曲率等。

3）焊接设备信息：包括要素为焊接电源品牌、焊接电源型号等。

4）焊接信息：包含的要素为焊接类型、焊接方法、焊接位置、焊材类型、焊材牌号、焊丝直径、摆动形式、摆动幅度、摆动速度、停留时间等。

5）保护气体：包含的要素为保护气体类型、混合气比例、纯度要求等。

6）焊接坡口信息：包含的要素为坡口形式、底部宽度、顶部宽度、衬垫形式（若有）、坡口角度、钝边量、错边量等。

7）施焊环境：包含的要素为周围环境的温度、湿度、风速等。

8）预热及道间温度：包含的要素为预热温度、预热方法、预热测温要求、层(道)间温度等。

9）焊接设备及电特性：包含的要素为焊接设备类型、电流种类、熔滴过渡形式等。

10）其他要求：包含的要素为焊工数量、焊枪摆动方式、单丝或多丝填充、背面清根方法、起弧和收弧要求、根焊与填充开始的时间间隔、焊丝干伸长、焊后保温、缓冷及热处理、焊接工艺参数等。

11）组队与焊接程序：包含但不限于根部清理、焊接材料、管口组队、焊口加热、焊接、无损检测、返修、其它(清理工具、施焊环境补充说明、切割焊口后的工艺选择)等。

d) 焊接接头检测报告数据

1）报告标识：包含的要素为报告编号、报告页码（第 X 页，共 X 页），唯一性标识等，便于归档和查询。

2）工件信息：包含的要素为产品名称、产品/工件编号、图号等。

3）焊缝信息：包含的要素为焊缝编号、焊缝类型（对接、角接等）、所属部件等。

4）工艺信息：包含的要素为焊接工艺规程（WPS）编号、焊材型号、规格等。

5）人员信息：包含的要素为操作人员（NDT 人员）姓名及资格证编号、审核人员签字等。

6）时间信息：包含的要素为检测日期、报告出具日期，记录检测时效等。

7）环境信息：包含的要素为环境温度、湿度（特别是对某些无损检测方法影响大）等。

8）依据标准：包含的要素为执行的技术标准（如：NB/T 47013.1）等。

9）检测设备：包含的要素为仪器型号、编号、校准有效期等。

10）检测方法：包含无损检测和破坏性检测。无损检测包括：射线检测（RT）、超声检测（UT）、磁粉检测（MT）、渗透检测（PT）、涡流检测/TOFD/相控阵（PAUT）；破坏性检测包括：力学性能试验、宏观金相/微观金相、硬度试验等。

11）检测结论：明确检测的数量及检测结果，如：本部件共拍片 X 张，其中 I 级合格 X 张、II 级合格 X 张。不合格宜列出所有发现的缺陷及其性质、尺寸、位置、等级。

5.1.2 图像数据采集要求

a) 坡口图像数据、熔池图像数据

1）坡口位于坡口图像中心位置。

2）熔池区域位于图像中心位置。

3）图像分辨率要求：激光图像分辨率≥1920×1200 像素，熔池图像分辨≥1280*1024 像素。

4）像素位深≥8bits。

5）宜记录图像采集的元数据。

6）保存格式包括但不限于 png、jpg 等。

b) 坡口形式示意图像

1）坡口示意图像分辨率≥224×224 像素。

2）在坡口形式示意图中宜能清晰观测其坡口形式、接头形式、焊缝形式。

3）坡口形式示意图中宜清晰标注出坡口角度、面角度、接头根部、根部间隙、根部半径、钝边等。

4）保存格式包括但不限于 png、jpg 等。

c) 焊接接头成型图像

1）图像分辨率≥224×224 像素。

2）焊缝图像宜清晰可量化，带有标尺。

3）焊缝图像宜真实无干扰，表面清洁、均匀光、纯背景。

4）宜记录图像采集的元数据。

5）保存格式包括但不限于 png、jpg 等。

d) 焊接接头质量检测图像

1）图像分辨率≥224×224 像素。

2）检测图像清晰可量化。

3）保存格式包括但不限于 png、jpg 等。

5.1.3 视频数据采集要求

1）采集帧率≥30fps。

2）保存格式包括但不限于 .avi、.mp4 等。

3）宜记录视频采集的元数据。

5.1.4 音频数据采集要求

1）位深≥16bit。

2）采集频率≥48kHz。

3）保存格式包括但不限于 .wav 等。

4）宜记录音频采集的元数据。

5.2 数据质量控制

数据总体上宜遵循数据的规范性、一致性、实时性、准确性、完整性、可访问性的要求；数据质量的六个指标由 GB/T 36344—2018 定义并给出。

5.2.1 文本数据质量控制

a) 宜重点确保工艺参数、检测报告等文本的准确性、完整性、一致性等。

b) 必填字段验证（如焊接设备、焊接方法、材料型号/牌号、电流电压参数是否缺失等）。

c) 空值/占位符检测（如"NULL"、"待补充"等无效值等）。

d) 宜符合格式一致性，包含但不限于正则表达式验证（日期格式 YYYY-MM-DD、数值单位统一为 A/V/mm/s）枚举值校验（焊接方法限于 MIG/TIG/激光焊等预设列表）等。

e) 宜符合逻辑合理性，包含但不限于范围阈值检查（电流值是否在 50-500A 合理区间）参数关联性验证（如：板厚 > 5mm 时电流值不宜低于 100A）等。

f) 专业术语标准化。

g) 宜采用的分析工具包含但不限于：Python + Pandas、Elasticsearch、规则引擎（Drools）实现复杂逻辑验证等。

5.2.2 图像数据质量控制

a）宜重点确保坡口图像、熔池图像、焊接接头、缺陷影像的清晰度、标注准确性。

b）图像基础质量验证，包含但不限于：分辨率验证、模糊度检测。

c）内容有效性验证，包含但不限于：ROI 区域检查（例如：焊缝区域占比 ≥30% 图像面积）、遮挡/反光检测（过曝区域面积 ≤5%）。

d）宜采用的分析工具包含但不限于：OpenCV。

5.2.3 视频数据质量控制

a）宜重点确保焊接过程动态记录的连贯性与关键帧质量。

b）时序完整性验证，包括但不限于：帧率稳定性、关键动作覆盖度（起弧/收弧过程无中断）。

c）多模态对齐：音画同步检测、传感器时序对齐检测。

d）关键帧质量检测：包含但不限于：抽取关键工艺帧（送丝速度突变点）进行图像质量检查、运动模糊检测（Sobel 边缘梯度分析）。

e）元数据校验，包含但不限于：时间戳连续性、设备 ID 与焊接参数绑定。

f）宜采用的分析工具包含但不限于：FFmpeg（帧抽取/同步分析）、光流法（运动模糊检测）、时间序列数据库。

5.2.4 音频数据质量控制

a） 宜重点确保保障电弧声、异常音的有效采集与特征可辨识度。

b）信噪比（SNR）检测，车间环境噪声 ≤45dB（电弧声频段 2-10kHz 信噪比 >20dB）。

c）特征完整性检测，包含但不限于：有效音频时长（单段 ≥ 5s）、关键事件标记（爆裂声、断弧声人工标注）。

d）失真检测，包含但不限于削波失真（振幅持续饱和比例 ≤1%）、采样率一致性。

5.2.5 知识图谱数据质量控制

a）模式层质量控制，包含但不限于本体冲突检测、属性约束验证。

b）三元组质量控制，包含但不限于跨源一致性校验、实体对齐消歧、工艺链路完整性。

c）动态质量控制，包含但不限于时效性分析、变更传播验证。

5.3 数据命名规则

5.3.1 文本数据

a) 焊接领域相关书籍、标准、规范、文献等数据

对焊接领域的书籍、标准、规范、文献等数据，根据不同类型的文本，其命名规则有不同的要求，现对命名规则进行统一规定。书籍类命名规则说明见表 1。

表 1 书籍类命名规则说明

类型	命名规则	举例
书籍	作者 - 书名+出版社+版本+出版年份.pdf	刘某某 -《钨极惰性气体保护焊(TIG)工艺与实践》[机械工业][第 2 版][2019].pdf
标准	标准代号+顺序号-年号+全称.pdf	GB/T 42127-2022《智能制造 工业数据 采集规范》.pdf
规范	规范代号-年号+全称.pdf	GB50661-2011《钢结构焊接规范》.pdf
文献	作者_出版年份_关键词_期刊或类型.pdf	赵五_2022_钛合金-电子束-增材制造_材料工程.pdf

b) 焊接作业过程中产生的焊接工艺参数数据

项目编号_产品编号_焊缝编号_层号_道号_焊接方法_母材材质_开始时间.扩展名

c) 焊接工艺规程数据

WPS-编号-材质-板厚-位置-方法-焊接设备型号-日期

d) 焊接接头检测报告数据

工单号_检测标准_焊缝编号_检测方法_母材材质_检测日期_版本.扩展名。焊接接头检测报告数据字段说明见表 2。

表 2 焊接接头检测报告数据字段说明

字段	说明	示例
工单号	关联工单的唯一标识	AJ-2025-002

检测标准	依据的检测标准（国际/行业/企业/团标）	ISO 5817、AWS D1.1、GB/T 3323
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
检测方法	检测技术缩写（需行业通用）	UT（超声）、RT（射线）、PT（渗透）、MT（磁粉）
母材材质	母材材质	Q345B
检测日期	报告生成日期（年月日）	20240615、2024-06-15
版本	修订版本号（初版可省略）	v1、Rev2
扩展名	不同格式的文件	.pdf

5.3.2 图像数据

a) 坡口图像数据

项目编号_产品编号_焊缝编号_层号_道号_焊缝类型_开始时间.扩展名。坡口图像数据说明见表 3。

表 3 坡口图像数据说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊缝类型	坡口形式	对接接头、角接接头、T型接头、搭接接头等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.png、.tiff、 .jpg等

b) 熔池图像数据

项目编号_产品编号_焊缝编号_层号_道号_焊接方法_开始时间.扩展名。熔池图像数据说明见表 4。

表 4 熔池图像数据说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001

焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊接方法	焊接工艺所使用的方法	TIG、MIG、MAG等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.png、.tiff、 .jpg等

c) 坡口形式示意图像

坡口类型_坡口角度_坡口间隙_其他参数.扩展名。坡口形式示意图像说明见表 5。

表 5 坡口形式示意图像说明

字段	说明	示例
坡口类型	坡口的形状	I型坡口（无坡口）、V型、U型、X型等
坡口角度	坡口组对后的角度	45°
坡口间隙	坡口组对后底部间隙	2mm
其他参数	钝边、错边、半径等，单位需明确	2mm等
扩展名	不同格式的文件	.png、.tiff、.jpg等

d) 焊接接头成型图

项目编号_产品编号_焊缝编号_层号_道号_焊缝类型_焊接方法_开始时间.扩展名。焊接接头成型图说明见表 6。

表 6 焊接接头成型图说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊缝类型	坡口形式	对接接头、角接接头、T型接头、搭接接头等

焊接方法	所使用的焊接方法	TIG、MIG、MAG等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.png、.tiff、 .jpg等

e) 焊接接头检测图

工单号_检测标准_焊缝编号_检测方法_材料_检测日期_版本.扩展名。焊接接头检测图说明见表 7。

表 7 焊接接头检测图说明

字段	说明	示例
工单号	关联工单的唯一标识	AJ-2025-002
检测标准	依据的检测标准（国际/行业/企业）	ISO 5817、AWS D1.1、GB/T 3323
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
检测方法	检测技术缩写（需行业通用）	RT（射线）
材料/部件	母材材质或部件名称	Q345B
检测日期	报告生成日期（年月日）	20240615、2024-06-15
版本	修订版本号（初版可省略）	v1、Rev2
扩展名	不同格式的文件	.pdf

5.3.3 视频数据

a) 激光视频数据

项目编号_产品编号_焊缝编号_层号_道号_焊缝类型_开始时间.扩展名。激光视频数据说明见表 8。

表 8 激光视频数据说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊缝类型	坡口形式	对接接头、角接接头、T型接头、搭接接头等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.mp4、.avi等

b) 熔池视频数据

项目编号_产品编号_焊缝编号_层号_道号_焊接方法_开始时间.扩展名。熔池视频数据说明见表 9。

表 9 熔池视频数据说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊接方法	焊接工艺所使用的方法	TIG、MIG、MAG等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.mp4、.avi等

5.3.4 音频数据

a) 电弧声音数据

项目编号_产品编号_焊缝编号_层号_道号_焊接方法_开始时间.扩展名。电弧声音数据说明见表 10。

表 10 电弧声音数据说明

字段	说明	示例
项目编号	项目关联信息	BX123
产品编号	项目关联信息	X001
焊缝编号	焊缝的唯一标识（可含位置或图纸编号）	001、bx002
层号	焊接层数号	3
道号	多层多道焊接，焊接道数号	3
焊接方法	焊接工艺所使用的方法	TIG、MIG、MAG等
开始时间	焊接开始时间	2024-06-15 12:23:34
扩展名	不同格式的文件	.wav、.mp3、.flac等

6 数据处理

6.1 数据清洗

6.1.1 多源数据清洗总体规则

文本、图像、视频、音频等数据清洗宜满足如下总体规则：

a) 宜制定明确的流程，即制定详细的、步骤化的清洗流程文档，明确每个清洗步骤的顺序、参数、使用的工具、脚本、词典版本。

b) 宜采用自动化的脚本进行清洗，主要清洗步骤宜通过可复用的脚本（Python 为首选）实现自动化，脚本需有良好的注释和日志记录功能。

c) 所有清洗脚本、映射表、词典、停用词表宜进行严格的版本控制（如 Git）。

d) 宜进行清洗质量控制： 非法字符去除率、编码转换成功率、术语标准化覆盖率与准确率（抽样检查）、数据一致性检查（如单位表示、标准号格式）等，例如：对每个数据批次进行抽样（如 1-5%），人工检查清洗前后的样本，评估清洗效果和潜在问题（如是否误删关键参数、术语标准化是否正确）。

e) 宜对清洗规则、词典、映射表进行迭代优化，需根据清洗检查结果、下游模型训练/应用的反馈，定期评审和更新清洗规则、词典、映射表。

f) 宜进行领域专家审核： 关键步骤（术语标准化、清洗规则制定、抽样检查）必须有焊接领域专家参与审核确认。

6.1.2 文本数据清洗

文本数据清洗除满足总体规则外，还宜满足：

a) 移除噪声、错误、无关信息，保留焊接核心语义。

b) 统一文本格式、编码方式。

c) 移除纯装饰性或干扰性标点（如过多感叹号、星号、不规则符号），但保留小数点、连接符、数学与单位符号、句末标点、引号、括号等。

d) 宜转换为简体中文，将数字、字母、标点符号的全角字符转换为半角字符。

e) 宜统一采用标准化术语，将非标准单位表示转换为标准单位符号。

6.1.3 图像数据清洗

图像数据清洗除满足总体规则外，还宜满足：

a) 宜保证图像基础质量，包括删除完全模糊、过度曝光或欠曝光图像、 纯色/空白图像、重复图像、统一分辨率及色彩模式。

b) 宜对图像内容进行清洗，包括裁剪或删除与焊接无关的区域（如背景人物、设备无关部分）、去除水印、时间戳、企业 LOGO（若涉及隐私）、马赛克处理人脸、身份证等敏感信息、视频抽帧图像需去重，保留关键帧（如电弧稳定、熔池清晰、缺陷可见的帧）。

6.1.4 视频数据清洗

视频数据清洗除满足总体规则外，还宜满足：

a) 宜保证视频基础质量，包括删除全程模糊/过曝/黑屏视频、无焊接主体（如仅拍摄设备或环境）、重复视频（内容≥90%重叠）。

b) 宜对视频内容进行清洗，包括裁剪焊接前/后的空镜头、人员走动、设备调试等非焊接过程片段、

模糊化或马赛克处理人脸、工牌等。

6.1.5 音频数据清洗

音频数据清洗除满足总体规则外，还宜满足：

a) 宜确保音频基础质量，删除静音片段（持续≥2 秒且无焊接声）、纯噪声（无焊接特征频率）、重复音频（波形相似度≥95%）等，保证统一的采样率与位深、声道与格式。

b) 宜对音频内容进行清洗，包括 模糊化或删除人员对话（如“准备焊接”等指令需保留），替换为白噪声或静音。

6.2 数据预处理

6.2.1 多源数据预处理总体规则

文本、图像、视频、音频等数据预处理宜满足如下总体规则：

a) 宜确保所有模态数据满足焊接领域专业性、一致性和完整性要求。

b) 预处理流程宜支持自动化、模块化，适应工业级数据规模。

c) 宜确保安全合规：处理敏感信息（工艺参数、人脸、设备铭牌等），符合保密要求。

d) 宜对数据预处理质量进行评估。

e) 多模态数据间宜建立跨模态数据的时空关联关系，支持联合建模。

f) 宜记录所有预处理操作（如裁剪参数、术语替换记录），保留原始数据副本，具有可追溯性。

6.2.2 文本数据预处理

文本数据预处理除满足总体预处理规则外，还宜满足：

a) 宜对文本数据预处理质量进行评估，包括但不限于术语标准化准确率、参数缺失率等。

b) 宜使用支持焊接领域术语的分词工具，用焊接专业词典进行强制细分或合并，确保术语完整性。

6.2.3 图像数据预处理

图像数据预处理除满足总体预处理规则外，还宜满足：

a) 图像与对应文本描述（如工艺参数）、传感器数据（电流/电压）时间戳对齐。

b) 宜进行噪声抑制，降低烟尘、飞溅、反光导致的噪声。

c) 宜采用焊缝区域清晰度（Canny 边缘强度）等对图像数据处理质量进行评估。

6.2.4 视频数据预处理

视频数据预处理除满足总体预处理规则外，还宜满足：

a) 视频与同步的传感器数据（电流、电压）、音频（电弧声）、文本（工艺日志）时间对齐。

b) 宜按焊接阶段提取关键帧：起弧阶段（电弧初始稳定性）、稳态阶段（熔池形态）、缺陷出现帧（裂纹、气孔）。

c) 宜采用关键帧覆盖率、时间同步误差等对视频数据预处理进行质量评估。

6.2.5 音频数据预处理

音频数据预处理除满足总体预处理规则外，还宜满足：

a) 音频与同步的传感器数据（电流、电压）、视频关键帧等进行时间对齐。

b) 宜进行噪声抑制，抑制环境噪声（风机、机械振动）、人员对话（非焊接相关）、电磁干扰（电流杂音）。

c) 宜采用信噪比（SNR）等对音频数据预处理质量进行评估。

6.3 数据增强

6.3.1 多源数据增强总体原则

a) 增强后的数据必须符合焊接物理规律。

b) 宜记录增强方法及参数（如旋转角度、噪声强度），原始数据与增强数据需关联存储。

c) 增强后数据需通过质检（如缺陷形态合理、文本参数有效），避免引入错误样本。

d) 对于多模态对齐的数据，增强单模态数据时，宜同步调整关联模态数据。

6.3.2 文本数据增强

a) 文本数据增强的方法包括但不限于：同义词替换、参数扰动、语法结构变换、对抗样本生成。

b) 文本数据增强的工具包括但不限于：nlpaug（同义词替换）、自定义焊接术语扰动规则（YAML配置）。

6.3.3 图像数据增强

a) 图像数据增强方法包括但不限于：几何变换、色彩调整、噪声注入、生成合成数据。

b) 图像数据增强工具包括但不限于：OpenCV（几何变换）、Albumentations（色彩/噪声）、StyleGAN3（合成数据生成）。

6.3.4 视频数据增强

a) 视频数据增强方法包括但不限于：时序采样、动态噪声模拟、多视角合成。

b) 视频数据增强工具包含但不限于：FFmpeg（帧率调整）、PyTorch Video（动态噪声生成）。

6.3.5 音频数据增强

a) 音频数据增强方法包括但不限于：时频域变换、环境噪声混合、声学特征增强。

b) 音频数据增强工具包括但不限于：Librosa（时频变换）、Audiomentations（噪声注入）。

7 数据标注

数据标注总体上宜遵循 GB/T 42755-2023 中的要求。

7.1 数据标注规范

7.1.1 文本数据标注

a) 文本数据标注包括但不限于：命名实体识别、关系抽取、多模态关联。

b) 标注的文本数据集宜采用的格式包含但不限于：JSON 格式、TXT 格式。

7.1.2 图像数据标注

a) 图像数据标注包括但不限于：目标检测、语义分割、关键点检测。

b) 图像的文本数据集宜采用的格式包含但不限于：JSON 格式、TXT 格式。

7.1.3 视频数据标注

a) 视频数据标注包括但不限于：帧级标注、时间事件标注、多目标追踪。

b) 视频的文本数据集宜采用的格式包含但不限于：JSON 格式、TXT 格式。

7.1.4 音频数据标注

a) 音频数据标注包括但不限于：事件标注、声学特征标注、异常检测。

b) 音频的文本数据集宜采用的格式包含但不限于：JSON 等。

7.2 数据标注质量控制

数据标注质量控制总体上宜遵循 GB/T 42755-2023 中 6 标注任务执行中的要求。

7.2.1 文本数据标注质量控制

a) 宜进行术语一致性检查，包含但不限于术语缩写不规范、参数单位错误、材料牌号映射错误。

b) 宜进行逻辑矛盾检测，包含但不限于数值超限、时空冲突、缺陷-参数矛盾等。

c) 宜进行多模态关联检查，包含但不限于图像引用存在性、参数同步性、缺陷描述一致性。

7.2.2 图像数据标注质量控制

a) 宜进行几何精度检查，包含但不限于标注框 IoU、边缘贴合度、多视角一致性。

b) 宜进行语义合理性检查，包含但不限于缺陷误分类、伪影误标、工艺矛盾。

c) 宜进行多模态关联检查，包含但不限于文本-图像尺寸、参数-熔池形态、时间-缺陷演化。

7.2.3 视频数据标注质量控制

a) 宜进行时序精度检查，包含但不限于关键帧同步误差、事件持续时间、轨迹平滑度。

b) 宜进行动态合理性检查，包含但不限于逆向运动、瞬态事件漏标、工艺-画面失配。

c) 宜进行多模态关联检查，包含但不限于音频-视频事件、文本-动态过程、传感器-视觉参数。

7.2.4 音频数据标注质量控制

a) 宜进行时域精度检查，包含但不限于事件同步误差、持续时间合规性、标注密度。

b) 宜进行频域特征检查，包含但不限于频段误标、能量分布异常、谐波缺失。

c) 宜进行多模态关联检查，包含但不限于视频-音频事件、文本-声学描述、传感器-频谱特征。

8 数据隐私与安全

数据隐私与安全总体上宜满足 GB/T 37973—2019 信息安全技术 大数据安全管理指南 中第 8 大数据活动安全要求中的要求。

8.1 数据安全传输要求

a) 数据在传输过程中宜满足机密性、完整性、可用性、审计追溯等要求。

b) 数据按照安全等级进行划分，不同的数据安全等级对应不同的网络传输要求。见表 11。

表 11 网络传输要求

安全等级	数据类型	网络传输要求

L1（最高）	军工/航天焊接参数、专利工艺数据	国密SM4加密 + 量子密钥分发 + 单向物理隔离
L2（高）	工业现场缺陷图谱、设备传感器数据、实时作业数据	TLS 1.3 + AES-256 + 双向身份认证
L3（中）	公开研究用焊接文本、脱敏视频	HTTPS + 动态令牌验证
L4（基础）	公开标准文档、教学演示数据	基础VPN通道 + MD5校验

c) 数据宜加密后传输，不同模态的数据对应的加密算法见表 12。

表 12 加密算法

数据类型	加密算法	密钥管理
文本参数	SM4（国密）/AES-256-GCM	硬件加密机（HSM）托管，每8小时轮换
图像/视频	H.265+SEI帧级加密	基于区块链的分布式密钥分发
音频流	OPUS编码+SRTP封装	动态会话密钥（DTLS 1.2握手）

8.2 数据隐私保护

a) 分级动态脱敏:根据数据敏感等级（P1-P4）实施差异化保护。

b) 多模态去标识化:视频模糊人脸与工装标识但保留焊枪操作特征，音频剥离声纹但保留工艺声学特征，文本泛化材料牌号（如"304→3XX"），确保数据可用性同时消除个体关联性。

c) 工艺感知的差分隐私：基于材料类型（如碳钢/钛合金）注入自适应噪声（±0.2%~0.5%），在保护核心参数的同时满足模型训练精度需求。

d) 全生命周期可信控制。

e) 合规应急双保障:建立三级隐私事件响应机制（核心参数泄露数据自动锁定），建立审计日志追踪。

8.3 数据安全使用方法

a) 分级权限管控

基于角色动态授权（工程师/质检员/访客），工艺参数需双因子认证，军工数据限制物理隔离环境使用，操作日志全链路审计。

b) 动态水印追踪

所有数据加载隐形水印（含操作员 ID/时间戳），任何泄露可追溯至具体环节，支持实时阻断外传行为。

c) 应急熔断机制

异常访问触发三级响应（预警/限流/熔断），核心参数泄露时自动擦除焊接设备缓存数据，为司法取证提供借鉴。