ALIBABA CLOUD

Alibaba Cloud

Machine Learning Platform for AI

Data Preparation

Document Version: 20211115

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	**Danger:** Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	**Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

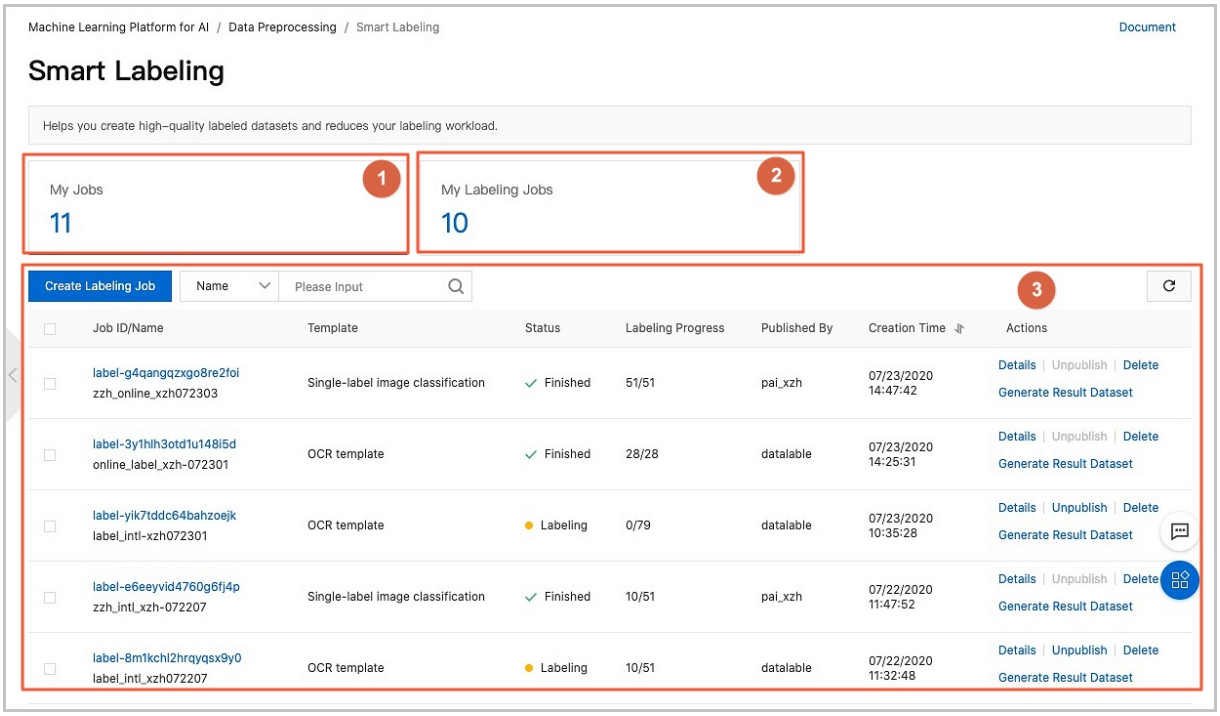
Table of Contents

1.Smart Labeling

1.1. Overview

You can view labeling jobs assigned to different workers on the Smart Labeling page.

On the Smart Labeling page, you can view **My Jobs**, **My Labeling Jobs**, and **Job List**.



Callout	Section	Description
①	**My Jobs**	You can perform the following operations on the **My Jobs** tab: • View labeling jobs that you have created as an administrator. • Click the ID of a labeling job in the **Job ID/Name** column. On the page that appears, view details about the job or modify the job. **Note** After you start a labeling job, **you are not allowed to modify the labeling policy** or **add labels**. • After you click **Generate Result Dataset**, the system generates a result dataset in the specified Object Storage Service (OSS) path. The result datasets generated in this path do not overwrite each other.
②	**My Labeling Jobs**	You can view the labeling jobs that are assigned to you and check the progress of these jobs.
③	**Job List**	Displays information about labeling jobs, including **Job ID/Name**, **Template**, **Status**, **Labeling Progress**, **Published By**, **Creation Time**, and **Actions**.

1.2. Data labeling templates

1.2.1. Data labeling templates

Machine Learning Platform for AI provides the following templates: object detection, semantic segmentation, comprehensive image annotation, Optical Character Recognition (OCR), single-label image classification, and multi-label image classification. When you create a labeling job, select a template that meets your requirements.

Object detection

Object detection is used to locate a specific object in an image. The rectangle selection tool is commonly used.

- Scenarios

  Vehicle detection, passenger detection, and image search.

- Data schema
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

    ```
    {"data":{"picUrl":"oss://****/pics/fruit/apple-1.jpg"}}
    ...
    ```

  - Output data

    Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

    ```
    {
      "data": {
        "picUrl": "oss://****/pics/fruit/apple-1.jpg"
      },
      "label-****(Labeling job ID)": {
        "results": [{
          "id": "oFVs8rkZ5E2PiMRmO****",
          "type": "rectangleLabel",
          "value": {
            "rotation": 0,
            "x": 39,
            "width": 353,
            "y": 12,
            "height": 292
          },
          "labels": ["Apple"]
        }]
      }
    }
    ```

Semantic segmentation

Semantic segmentation is used to recognize an object in an image and retrieve the coordinates of the object (by scanning all pixels of the object). The commonly used tools are the polygon selection tool, brush tool, and superpixel tool.

- Scenarios

  Autonomous driving, facial expression recognition, and apparel classification.

- Data schema
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

    ```
    {"data":{"picUrl":"oss://****/pics/fruit/apple-1.jpg"}}
    ...
    ```

  - Output data

    Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

```
{
  "data": {
    "picUrl": "oss://****/pics/fruit/apple-1.jpg"
  },
  "label-****(Labeling job ID)": {
    "results": [{
      "id": "dnWsxyJIAmydbBhtJ****",
      "type": "polygonLabel",
      "value": {
        "points": [
          [110, 46],
          [146, 22],
          [203, 10],
          [229, 11],
          [262, 23],
          [289, 38],
          [320, 66],
          [346, 97],
          [361, 149],
          [351, 197],
          [335, 231],
          [302, 265],
          [259, 288],
          [215, 297],
          [190, 300],
          [162, 298],
          [139, 290],
          [119, 285],
          [98, 267],
          [87, 257],
          [75, 243],
          [67, 226],
          [52, 196],
          [48, 168],
          [48, 145],
          [54, 120],
          [63, 93],
          [76, 74]
        ]
      },
      "labels": ["Apple"]
    }]
  }
}
```


Comprehensive image annotation

Comprehensive image annotation is used to match the content of the input images against a set of labels. This template allows you to use all image labeling tools.

- Scenarios

  Autonomous driving, content moderation, and content recoginition.

- Data schema

- Input data

  Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

  ```
  {"data":{"picUrl":"oss://****/pics/fruit/apple-10.jpg"}}
  ```

- Output data

  Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

  ```
  {
    "data": {
      "picUrl": "oss://****/pics/fruit/apple-10.jpg"
    },
    "label-****(Labeling job ID)": {
      "results": [{
        "data": [{
          "id":"Znyumd-****",
          "type":"image/rectangleLabel",
          "value":{
            "rotation":0,
            "x":40.68320610687023,
            "width":327.52035623409665,
            "y":5.762467474590647,
            "height":296.68117192104745
          },
          "labelColor":"#72bf7d",
          "labels":["Ripe apple"]
        }],
        "id":"44****",
        "type" : "image",
      }]
    }
  }
  ```

OCR

OCR is used to extract text from input images, and then classify the images based on the text.

- Scenarios

  Identity card, ticket, license plate, and bank card recognition.

- Data schema
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

    ```
    {"data":{"picUrl":"oss://****/img/ocr_card/img0.jpeg"}}
    ```

- Output data

  Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

```
{
  "data": {
    "picUrl": "oss://****/img/ocr_card/img0.jpeg"
  },
  "label-****(Labeling job ID)": {
    "results": [{
      "data": [{
        "direction_of_picture":"downward",
        "type":"ocr/meta"
      },
      {
        "id": "Y4ZFoC-****",
        "direction_of_text": "downward",
        "text": "Alibaba Cloud Intelligence",
        "type": "ocr/polygonLabel",
        "value": {
          "points": [[325.08789110183716,397.47582054138184]]
        },
        "labelColor": "#67bd3a",
        "labels": "Enterprise"
      }],
      "id":"24****",
      "type":"ocr"
    }]
  }
}
```

Single-label image classification

Single-label image classification is used to find a label from a set of labels to match the content of an input image, and then attach the label to the image.

- Scenarios

  Photo classification, image recognition, and image search.

- Data schema
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

```
{"data":{"picUrl":"oss://****/img/ocr_card/img0.jpeg"}}
```

- Output data

  Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

```
{
  "data": {
    "picUrl": "oss://****/img/ocr_card/img0.jpeg"
  },
  "label-****(Labeling job ID)": {
    "results": [{
      "data": [{
        "data":"red",
        "id":"33****",
        "type":"survey/value"
      }],
      "id":"33****",
      "type":"survey"
    }]
  }
}
```

Multi-label image classification

Multi-label image classification is used to find multiple labels from a set of labels to match the content of an input image, and then attach the labels to the image.

- Scenarios

  Content recommendation, advertising, and image search.

- Data schema
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the picUrl field.

```
{"data":{"picUrl":"oss://****/img/ocr_card/img0.jpeg"}}
```

- Each row in the **manifest** file contains a topic and the corresponding labeling result. The following is an example of the JSON string in each row:

```
{
  "data": {
    "picUrl": "oss://****/img/ocr_card/img0.jpeg"
  },
  "label-****(Labeling job ID)": {
    "results": [{
      "data": [{
        "data":["red","more","green"],
        "id":"33*****",
        "type":"survey/multivalue"
      }],
      "id":"33****",
      "type":"survey"
    }]
  }
}
```

1.2.2. Labeling templates for videos

Machine Learning Platform for AI provides the video classification and object marking templates. When you create a labeling job, select a template that meets your requirements.

Video classification

Video classification is used to find one or more labels from a set of labels to match the content of an input video and attach the labels to the video. The template supports single-label and multi-label video classification.

- Scenarios

  Video surveillance, live streaming recommendation, and short video recommendation.

- Data structure
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the videoUrl field.

```
{"data":{"videoUrl":"oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"}}
```

- Output data

  Each row in the **manifest** file contains a topic and the labeling result. The following code provides an example of the JSON string in each row:

```
{
 "data": {
  "videoUrl": "oss://xxxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/show.webm"
 },
 "label-bo0pc51ok7olwe****": {
  "results": [{
   "data": [{
    "data": ["Car promotion", "Promotion video"],
    "id": 165****,
    "type": "survey/multivalue"
   }],
   "id": "165****",
   "type": "survey"
  }]
 }
}
```

Object marking

Object marking is used to locate and mark specific objects in a frame or specific frames of a video. The rectangle selection tool and polygon selection tool are commonly used.

- Scenarios

  Autonomous driving, security surveillance, and video recommendation.

- Data structure
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the videoUrl field.

```
{"data":{"videoUrl":"oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"}}
```

  - Output data

    Each row in the **manifest** file contains a topic and the labeling results of one or more frames. The following code provides an example of the JSON string in each row:

```
{
 "data": {
  "videoUrl": "oss://xxxxxxxxx.oss-cn-shanghai-internal.aliyuncs.com/video/English.mp4"
 },
 "label-rv0ih5l409u9x9****": {
  "results": [{
   "data": [{
    "speed_of_play": 1,
    "duration": 300.733375,
    "width": 960,
    "number_of_frames": 9022,
    "type": "video/meta",
    "frame_rate": 30,
    "height": 540
   }, {
    "frames": {
     "frame-443": {
      "L6G-5Oc5STCSOuzKb****": {
       "rotation": 0,
       "x": 336.0483870967742,
       "width": 488.4677419354839,
       "y": 108.87096774193549,
       "height": 349.83870967741933
      }
     }
    },
    "type": "video/frames"
   }, {
    "frames": [443],
    "custom": {
     "Video type": "English learning",
     "Video name": "English learning",
     "Category": ["Education"]
    },
    "id": "L6G-5Oc5STCSOuzKb****",
    "type": "video/rectangleLabel",
    "value": "",
    "labelColor": "#ff7700"
   }],
   "id": "129****",
   "type": "video"
  }]
 }
}
```


1.2.3. Labeling templates for text

Machine Learning Platform for AI provides the text classification template. This topic describes the scenarios and data structure of the template.

Text classification

Text classification is used to find one or more labels from a set of labels to match the content of the input text and attach the labels to the text. The category template supports single-label and multi-label text classification.

- Scenarios

  News recommendation, knowledge management, and junk content filtering.

- Data structure
  - Input data

    Each row in the **manifest** file contains a topic. The topic must contain the content field.

    ```
    {"data":{"content":"Welcome to Machine Learning Platform for AI! "}}
    ```

  - Output data

    Each row in the **manifest** file contains a topic and the labeling result. The following code provides an example of the JSON string in each row:

    ```
    {
     "data": {
      "content": "Welcome to Machine Learning Platform for AI! "
     },
     "label-p9fypszvupm9bh****": {
      "results": [{
       "data": [{
        "data": "Positive content",
        "id": 112****,
        "type": "survey/value"
       }],
       "id": "112****",
       "type": "survey"
      }]
     }
    }
    ```

1.3. Create a labeling job

This topic describes how to create a labeling job.

Prerequisites

A dataset is registered. For more information about how to register a dataset, see Manage datasets.

Procedure

1. Log on to the Machine Learning Platform for AI console.
2. In the left-side navigation pane of the Machine Learning Platform for AI console, choose **Data Preprocessing > Smart Labeling**.
3. On the **Smart Labeling** page, click **Create Labeling Job**.
4. In the **Template** step, set the parameters and click **Next**.

Parameter		Description
Template		The template to be used in the labeling job. The system supports the following templates: ◦ **Image** ▪ **Object Detection** ▪ **Semantic Segmentation** ▪ **Image Comprehensive Label** ▪ **OCR** ▪ **Image Classification** ◦ **Text** **Text Classification** ◦ **Video** ▪ **Video Classification** ▪ **Object Annotation**
Labels		The labels that are used to classify images. This parameter takes effect only when you select **Object Detection**, **Semantic Segmentation**, or **Image Comprehensive Label** for the **Template** parameter.
Image and Text Orientation	**Label Image Orientation**	Specifies whether to annotates image orientation. This parameter takes effect only when you select **OCR** for the **Template** parameter.
	Flip Text	Specifies whether to annotates text orientation. If the text in an image is placed in the same direction as the image, you can turn off the switch. This parameter takes effect only when you select **OCR** for the **Template** parameter.
Text Type		The labels that are used to classify the text. This parameter takes effect only when you select **OCR** for the **Template** parameter.
Add Custom Label		The custom labels that are used to classify the text. This parameter takes effect only when you select **OCR** for the **Template** parameter.
Labeling Type		Valid values: **Single-label** and **Multi-label**. You can select the labeling type based on your needs. This parameter takes effect only when you select **Image Classification** for the **Template** parameter.
Labels		The labels that are used to classify images. Labels are displayed in different colors. This parameter takes effect only when you select **Image Classification** for the **Template** parameter.

5. In the **Basic Information** step, set the parameters and click **Next**.

Parameter	Description
Task Name	The name must be 1 to 30 characters in length and can contain underscores (_) and hyphens (-). It must start with a letter or a digit.

Parameter	Description
Description	The description must be 1 to 64 characters in length and can contain underscores (_) and hyphens (-). It must start with a letter or a digit.
Input Dataset	Select one or more datasets to create a labeling job. The datasets must correspond to the topic of the labeling job. If no dataset is available, click **Register Dataset** next to **Input Dataset** to register a dataset.
Output Dataset Path	The Object Storage Service (OSS) path to which the labeling results are stored. When you handle a labeling job, every time you click **Generate Result Dataset**, a result dataset is generated in the specified OSS path. The dataset contains the results of all topics that you have completed.

6. In the **Labeling Policy** step, set the parameters and click **Submit**.

Parameter	Description
Dispatch Policy	The default dispatch policy is **Number of topics collected by a worker each time** and cannot be changed.
Topics per Collection	The number of topics collected by each worker each time. **Note** The value of the **Topics per Collection** parameter can be smaller than the total number of topics divided by the total number of workers. This allows workers with high efficiency to collect more topics and improves the overall efficiency of data labeling.
Add Worker	You can specify one or more workers. You can select both Alibaba Cloud accounts and RAM users.

1.4. Label images

This topic describes how to label images.

Prerequisites

Create a labeling job or contact the administrator to assign a labeling job to you. For more information, see Create a labeling job.

Procedure

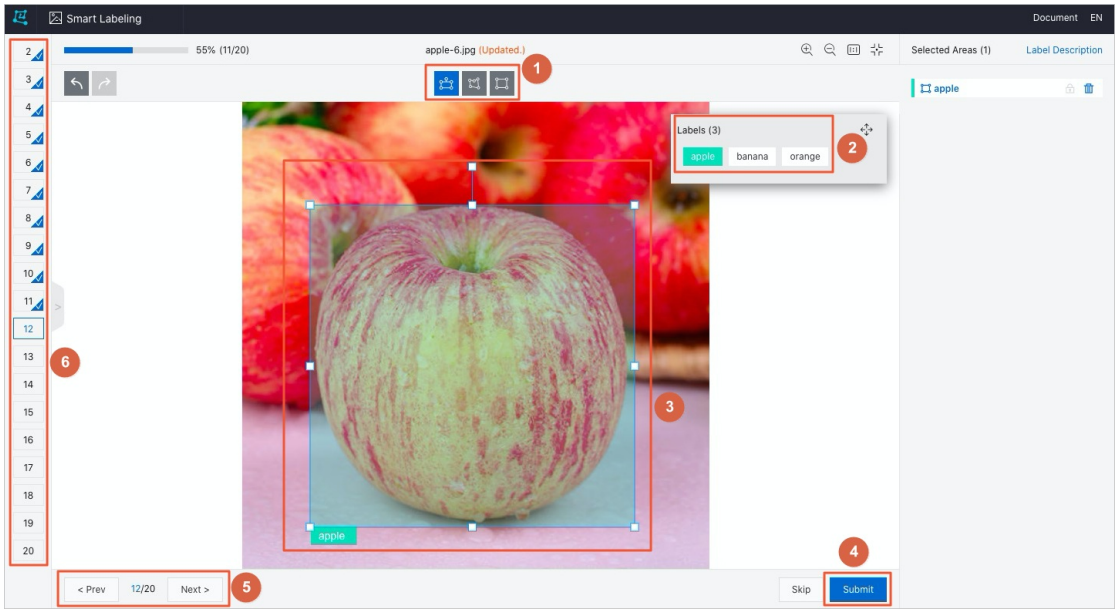
1. Navigate to the Smart Labeling page.
   i. Log on to the Machine Learning Platform for AI console.
   ii. In the left-side navigation pane, choose **Data Preprocessing > Smart Labeling** to navigate to the **Smart Labeling** page.
   iii. Click the **My Labeling Jobs** tab.
   iv. In the job list, find your labeling job and click **Start** in the **Actions** column.
2. Label images.

i. On the labeling page, click the icon.

ii. In the **Labels** pane, select a label.

> **Note** The system attaches the selected label to each image unless you select another label.

iii. Use the selection tool to select an area in the image.



iv. (Optional)If you do not need to select a specific area, click **Skip**.

v. Click **Submit**.

vi. You can use one of the following methods to browse and complete topics.

- Click **Prev** or **Next** at the bottom of the labeling page.
- Click the thumbnails of the topics in the left-side pane.

1.5. Shortcut keys

This topic describes how to use shortcut keys when you label images.

Function	Shortcut key
Use the rectangle selection tool	R
Use the polygon selection tool	P
Switch a label	1, 2, 3, 4, 5, 6, 7, 8, 9 (only for the first 9 labels)
Skip the current topic	S
Submit the current topic	Enter
Return to the previous topic	Left

Function	Shortcut key
Go to the next topic	Right

Note You can move the pointer over a button to check the shortcut key.

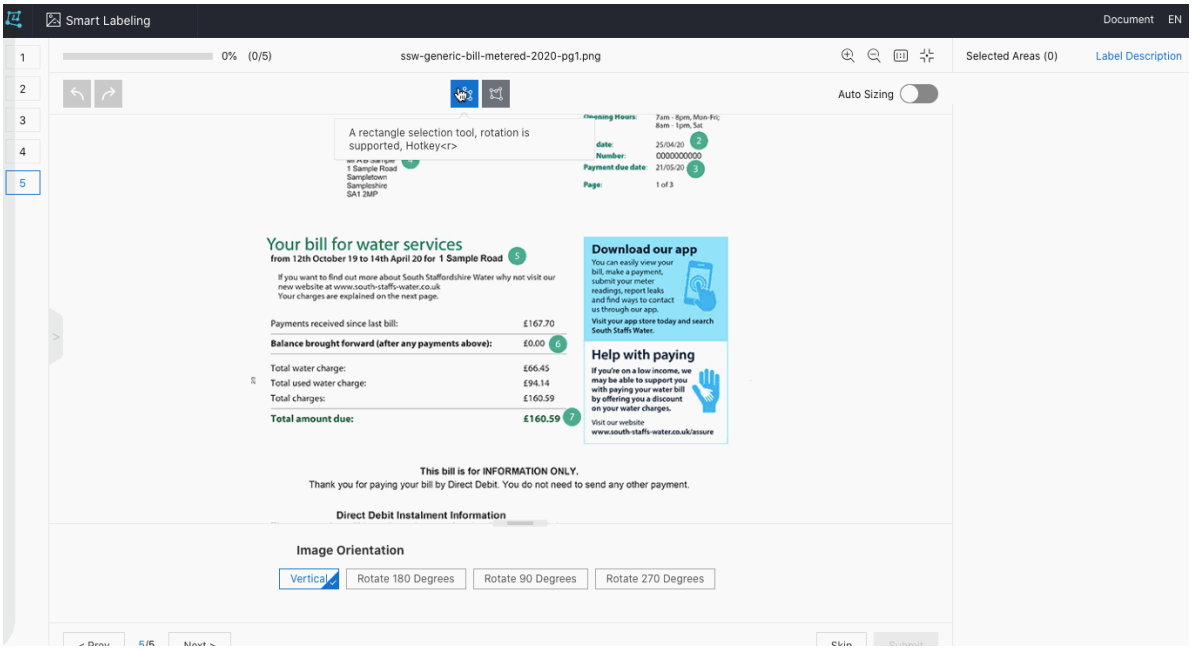
1.6. Smart text box fitting and character recognition

Traditional tools for labeling text in images require that you manually adjust the box to fit the text. This is characterized by low efficiency, and can easily bring about labeling errors when the text is overly dense. Smart labeling of Machine Learning Platform for AI supports text box fitting, which can be used to automatically adjust skewed, distorted, and congested text boxes. Smart labeling also supports character recognition, which can automatically recognize content in a selected area and then convert it to text. This enables you to complete labeling jobs efficiently.

Smart text box fitting

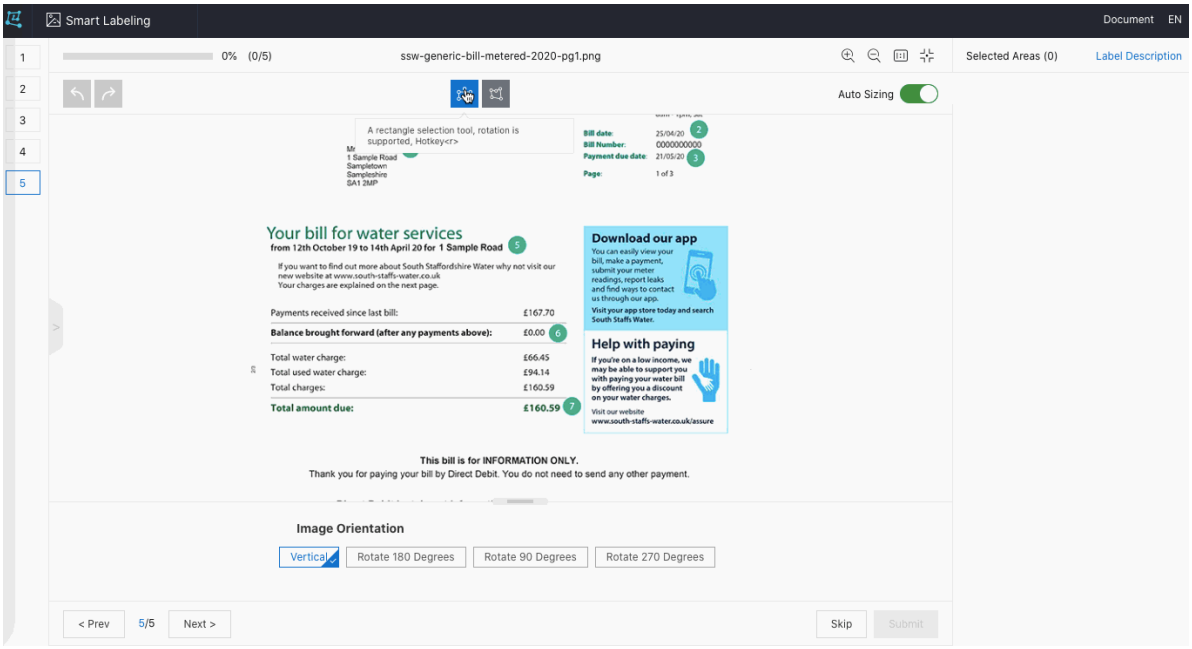
- Manual labeling

  Manual labeling is marked by low efficiency and cannot control the size of text boxes.



- Smart labeling

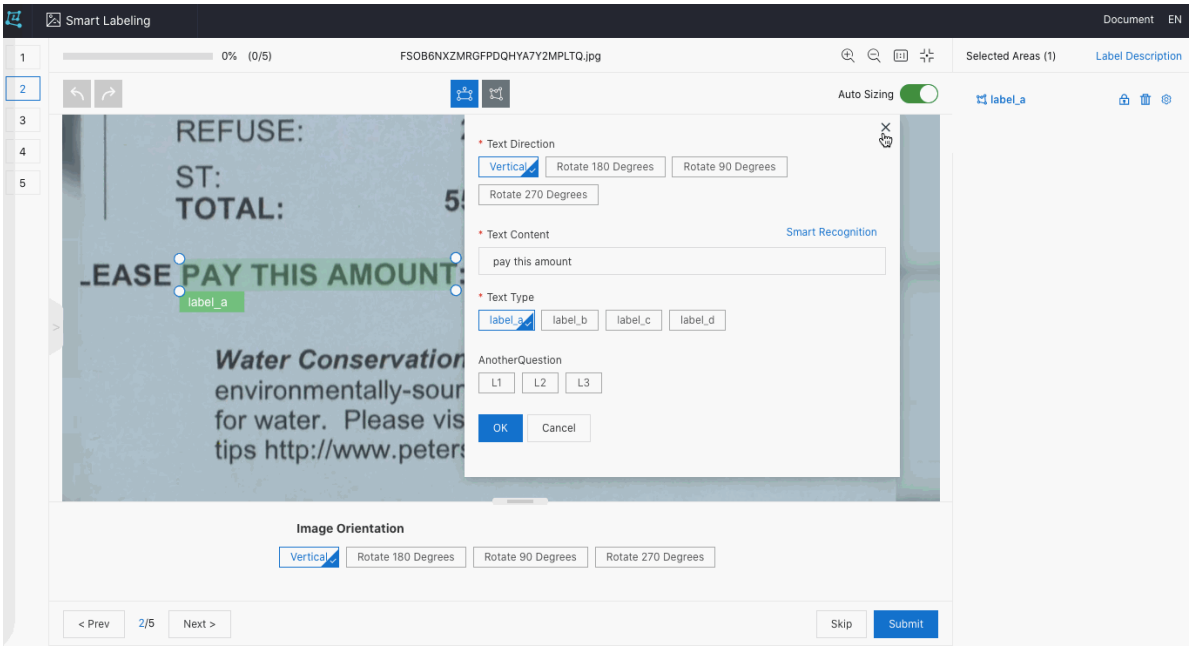
Text boxes can be scaled to fit the text by using the smart labeling feature. This allows you to improve the efficiency of data labeling.



Smart text recognition

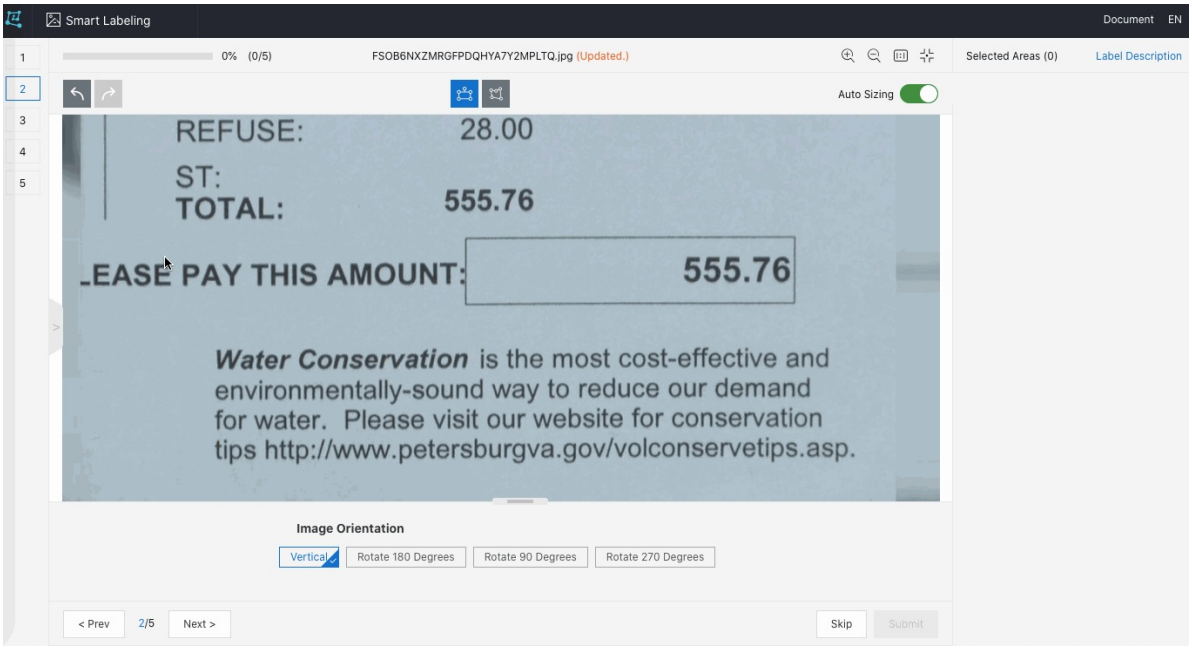
- Manual labeling

  After you select the content, you must enter the text, which leads to low efficiency.



- Smart labeling

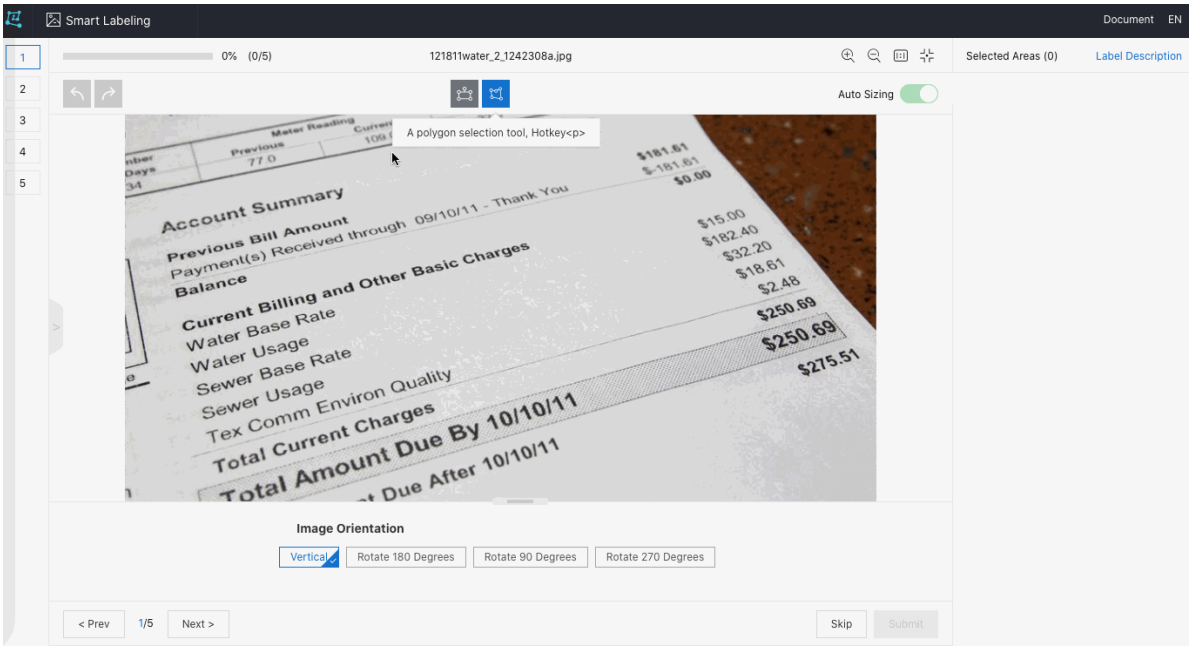
Select the text content, and then click **Smart Recognition**. The system can automatically recognize the text in the box and generate corresponding characters.



Skewed text labeling

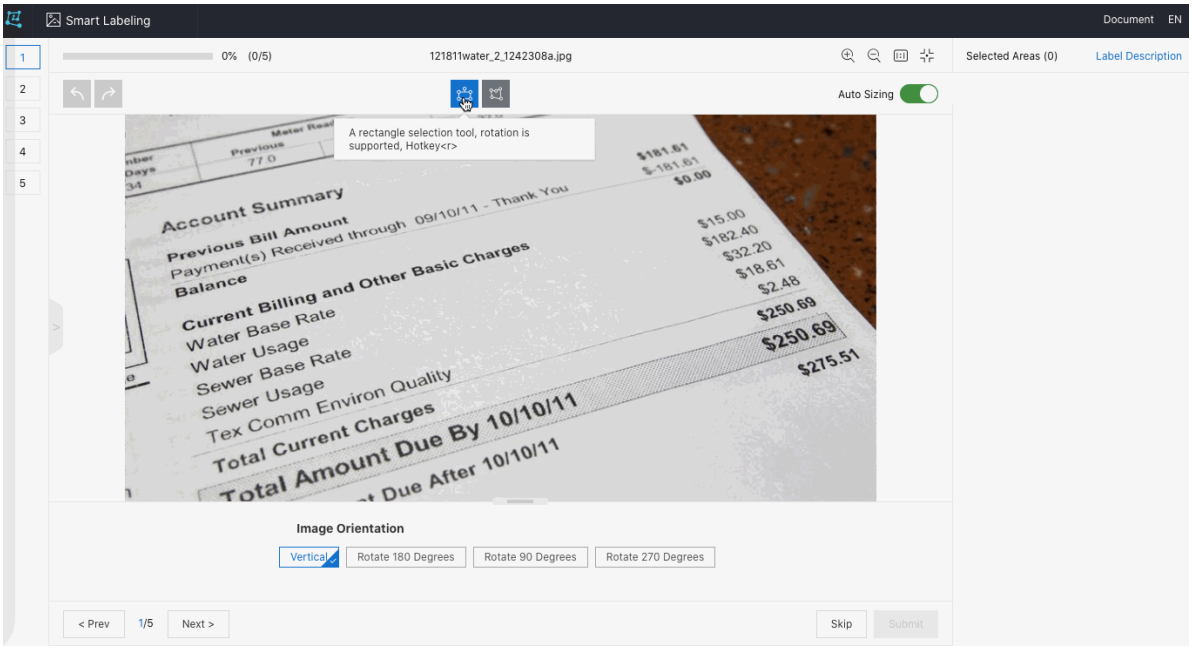
- Manual labeling

  Polygon selection tools are used for labeling, which can be inefficient.



- Smart labeling

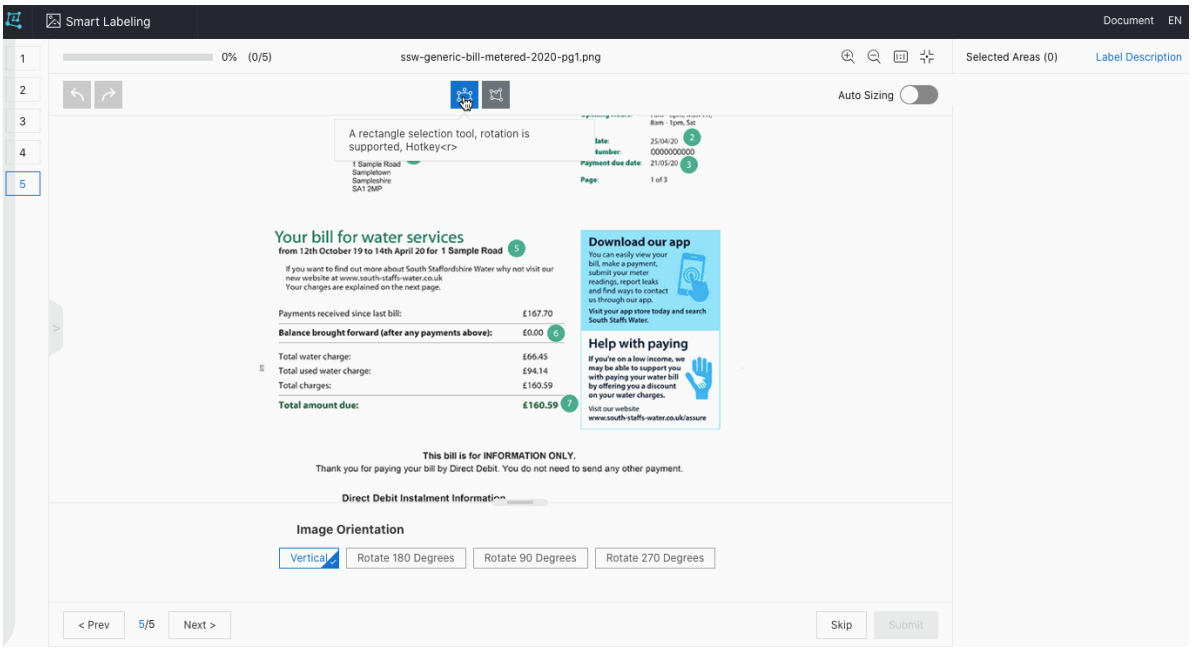
You only need to use the rectangle selection tool to select the skewed text area. The system can automatically identify the text range and fit the text.



Dense text labeling

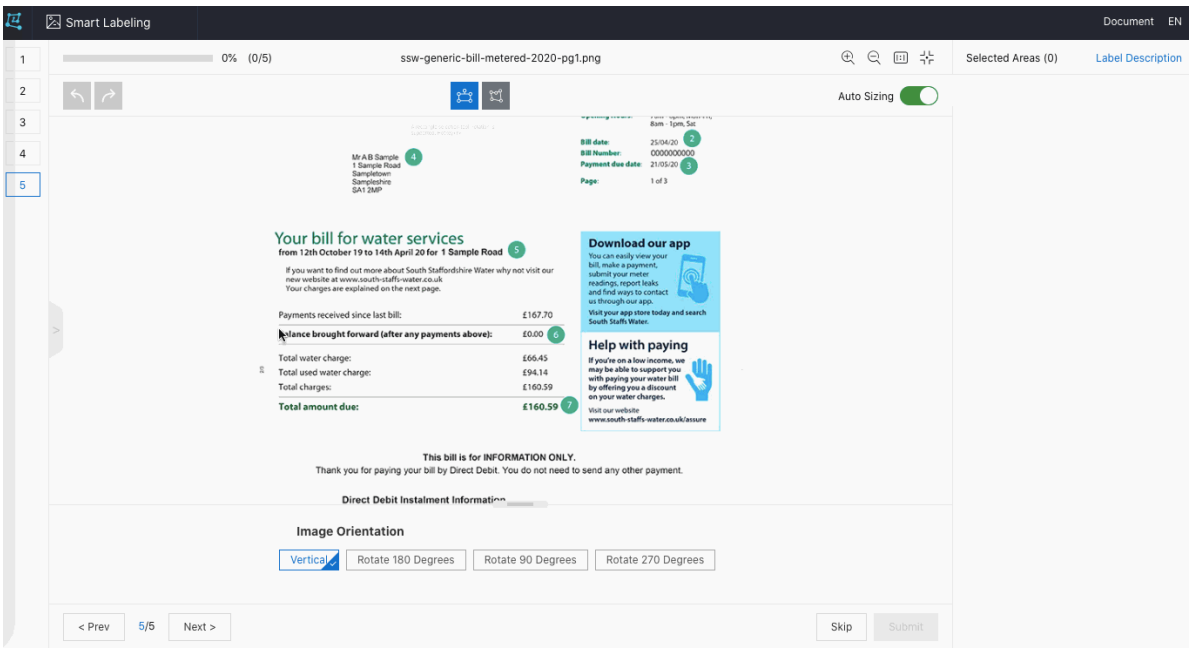
- Manual labeling

  When manual labeling is used in dense text, other text around the target text can be selected in the box, which may cause labeling errors.



- Smart labeling

To perform dense text labeling, you only need to use the rectangle selection tool to select the area that contains the dense text. The system can automatically fit the target text, even if undesired text is selected by mistake.



2.Use cases