# 阿里云 内容安全

用户指南

文档版本:20180911



# 法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读 或使用本文档,您的阅读或使用行为将被视为对本声明全部内容的认可。

- 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档,且仅能用于自身的合法 合规的业务活动。本文档的内容视为阿里云的保密信息,您应当严格遵守保密义务;未经阿里云 事先书面同意,您不得向任何第三方披露本手册内容或提供给任何第三方使用。
- 未经阿里云事先书面许可,任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分 或全部,不得以任何方式或途径进行传播和宣传。
- 由于产品版本升级、调整或其他原因,本文档内容有可能变更。阿里云保留在没有任何通知或者 提示下对本文档的内容进行修改的权利,并在阿里云授权通道中不时发布更新后的用户文档。您 应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
- 4. 本文档仅作为用户使用阿里云产品及服务的参考性指引,阿里云以产品及服务的"现状"、"有缺陷"和"当前功能"的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引,但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的,阿里云不承担任何法律责任。在任何情况下,阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害,包括用户使用或信赖本文档而遭受的利润损失,承担责任(即使阿里云已被告知该等损失的可能性)。
- 5. 阿里云网站上所有内容,包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站 画面的安排、网页设计,均由阿里云和/或其关联公司依法拥有其知识产权,包括但不限于商标 权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意,任何人不得擅自使 用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此 外,未经阿里云事先书面同意,任何人不得为了任何营销、广告、促销或其他目的使用、公布或 复制阿里云的名称(包括但不限于单独为或以组合形式包含"阿里云"、Aliyun"、"万网"等阿里云 和/或其关联公司品牌,上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或 服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联 公司)。
- 6. 如若发现本文档存在任何错误,请与阿里云取得直接联系。

# 通用约定

格式	说明	样例
•	该类警示信息将导致系统重大变更甚至 故障,或者导致人身伤害等结果。	禁止: 重置操作将丢失用户配置数据。
A	该类警示信息可能导致系统重大变更甚 至故障,或者导致人身伤害等结果。	<ul> <li>▲ 警告:</li> <li>重启操作将导致业务中断,恢复业务所需</li> <li>时间约10分钟。</li> </ul>
	用于补充说明、最佳实践、窍门等,不是用户必须了解的内容。	<b>送</b> 说明: 您也可以通过按 <b>Ctrl + A</b> 选中全部文件。
>	多级菜单递进。	设置 > 网络 > 设置网络类型
粗体	表示按键、菜单、页面名称等UI元素。	单击 确定。
courier 字体	命令。	执行 cd /d C:/windows 命令,进 入Windows系统文件夹。
斜体	表示参数、变量。	bae log listinstanceid Instance_ID
[]或者[a b]	表示可选项,至多选择一个。	ipconfig[-all/-t]
{}或者{a b}	表示必选项,至多选择一个。	<pre>swich {stand   slave}</pre>

# 目录

汥	全律声明	I
通	···· / / // 〔用约定	I
1	使用限制	
2	站点检测	2
3	<b>OSS</b> 违规检测	
4	内容检测API	
	4.1 授权访问MTS服务	
	4.2 数据统计	21
	4.3 数据回流	
	4.4 自定义图库	26
	4.5 自定义词库	

# 1 使用限制

本文描述了您在使用内容安全时应该注意的产品功能限制。

### 站点检测

站点检测的对象是您的网站上的网页和图片,以URL数量进行计数。在单个网站的一个检测周期内,站点检测支持的最大检测容量为10万个URL。

### OSS违规检测

- OSS违规检测只向阿里云OSS用户提供服务。使用该服务前,您需要在内容安全控制台,通过RAM,授权内容服务读取OSS Bucket的权限(该操作在您首次登录控制台时,可以一键完成)。
- 增量图片支持自动检测,但需要您在内容安全设置中添加具体的 Bucket;这样,内容安全才会 每日帮助您检测该 Bucket 内实时增加的图片。
- 存量图片不支持自动检测,如果您想要对存量图片进行检测,您需要手动选择 Bucket 和时间范围,进行存量扫描。

# 2 站点检测

站点检测服务定期检查您的网站首页和全站内容,及时发现您的网站在内容安全方面可能存在的风险(如首页篡改,挂马暗链,色情低俗,涉政暴恐等),并向您展示违规内容的具体地址,帮助您查看和修复。您可以设置消息通知,选择邮件、短信、站内信的方式,获取实时的站点首页风险提醒。

功能描述

购买站点检测实例后,您需要将实例绑定到您的站点,添加要检测的网站域名和首页地址,设定首页和全站检测的频率,并完成网站鉴权。完成设置后,系统将定期按照您设定的频率对首页和全站内容(包含网页源码,文本和图片)进行检测。如果发现有风险,会按照您设定的消息接收方式通知您,您也可以登录产品控制台查看检测结果。

在对一个站点进行全站检测时,一个站点检测实例在一个检测周期内,支持的最大检测限制为:网页数和图片数合计不超过10万条/张。

站点检测提供以下功能:

• 首页监测

定期对您网站的首页进行监测,展示最近一次的检查结果,涵盖首页篡改,挂马暗链,色情低俗,涉政暴恐风险;提供源码,文本,图片三类呈现方式,供您参照和整改。

• 全站检测

定期对您网站域名下的网页进行自动化全站内容检测,展示最近一次的检查结果,涵盖挂马暗链,色情低俗,涉政暴恐风险;提供源码,文本,图片三类呈现方式,供您参照和整改。

站点检测支持以下设置:

• 设置首页防篡改基准

通过算法对比网页实时状态和您预设基准状态对比,判断是否为非法篡改。

• 添加重点监控URL

为了确保全站检测时重要页面不会遗漏,建议您添加网站重点监控URL。最多支持添加5,000条 重点监控URL。

• 自定义词库和图库

在使用站点检测服务时,您可以添加自定义关键词进行黑名单防控;添加的关键词会在15分钟内 生效,关键词只支持UTF-8格式。在使用站点检测进行鉴黄、暴恐等图像服务时,您可以添加自 定义图片进行黑名单/白名单防控;添加的图片会在15分钟内生效。

关于该功能的更多介绍,请参考文本反垃圾API。

#### 前提条件

已购买站点检测实例。

购买方法请参考购买站点检测实例。

操作流程

在购买站点检测实例后,您需要将实例绑定到待检测的站点,为站点启用检测服务。然后,您可以 在控制台查看首页检测和全站检测的结果。

如果您想进一步保障检测效果,您可以设置首页防篡改基准和重点监控URL,或者自定义词库,自定义图库。

您也可以使用消息通知设置,设置风险通知方式,开启/关闭首页风险实时通知。

#### 启用站点检测

参照以下步骤,为您的站点启用站点检测服务:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 站点检测页面。
- 3. 选择一个有效的,处于未绑定状态的实例,单击其操作列下的绑定站点。

设置						
消息通知	站点检测		OSS 违规检测	内容检测 API		
+ 购买实例	さ 刷新					
状态	域名	协议	检测频率	检测实例	使用有效期	操作
未绑定				cdisitecheck-cn-0pp0kavsp005	2018-04-13 至 2018-10-14	绑定站点

4. 在绑定站点对话框中,填写网站信息和检测频率。配置说明如下:

绑定站点		×
转度协议:	HTTP O HTTPS	
站市城名:	bb/bqbq01.com	
	如果站台包含多个二级地名,建议输入顶级地名,如 abc.com	
新以首页地址 ③	http://bb.logbq01.com	
重页检测间隔。	1 小时	
全站经网络车〇	● 低:7天1次 ○ 窟:1天1次	

配置	说明
站点协议	勾选HTTP或者HTTPS。如果您的网站HTTP和HTTPS分别对应不同的 内容,而且内容差异较大,建议您绑定2个不同的实例。
站点域名	填写您站点的域名,填写时不要包含http://或者https://。如果 您的网站有多个频道子域名,建议您在这里填写根域名。
	说明:         假设您的网站         有www.domain.com, news.domain.com, sports.domain.com等         多个频道内容需要检测,建议您在这里输入domain.com。如         果您只想检测news频道的内容news.domain.com, 您可以输入news.domain.com。
站点首页地址	填写完整的站点首页网址。输入的网址必须在您要绑定的域名下。
首页监测间隔	设置每隔多少小时,访问您的网站首页进行一次检测。默认为1个小时。
全站检测频率	设置执行全站检测的频率,7天1次或者1天1次。默认为7天1次。 说明: 站点检测频率越高,检测占用的带宽及产生的带宽费用也越多。如果 您的网站内容比较多,且网站带宽不足的话,过高的检测频率可能影 响您网站的正常访问速度。如果您不希望影响网站性能,建议您配置 较低的检测频率。

5. 单击下一步,进行站点验证,证明您对站点的管理权,防止未经授权的检测。

在验证站点对话框,选择验证方式,参照对话框中的验证说明完成相应操作,然后单击立即验证。如果您暂时不方便进行验证,您可以单击稍后验证,保存当前已输入的数据。支持的验证方式包括以下四种:

验证站点		×
書画编辑正方式,验证你对 bb.bqbq01.com 的结束所有权;		
<ul> <li>阿里云照户检证</li> </ul>		
① 主机文件程道		
○ CNAME 被名验证		
○ 网络黄页 HTML 标签验证		
同里云账户验证 2日前登费的问题云际今资产下有比较名		
		_
解集發達	上一步	30882

- 阿里云账户验证:验证待检测站点(域名)是否在您当前登录的阿里云账号的资产下。
- 主机文件验证:按要求在域名对应主机的根目录下生成相应的文件进行验证。
- CNAME验证:按要求在待检测域名的解析记录中增加指定的CNAME记录进行验证。关于添加CNAME记录的操作,请参考添加CNAME记录。
- 网站首页HTML标签验证:按要求修改网站首页HTML源文件进行验证。
- 7. 验证通过后,完成站点绑定和检测设置,目标实例自动开始检测。

回到设置 > 站点检测页面,选择已绑定站点的实例,在其操作选项中,您可以执行以下操作:

- 暂停/启动检测:如果您不希望在当前时间执行检测,您可以暂停检测;已暂停的检测,通过启动检测可以恢复。
- 编辑站点:修改实例绑定的站点和检测频率信息。



如果您修改了站点或首页地址,需要重新验证。

- 重新验证:如果您的验证失效或在步骤6中选择了稍后验证,您可以重新验证对站点的管理权。
- 解除绑定:如果您不希望继续向已绑定的站点提供检测服务,您可以解除绑定。

### ■ 说明:

解除绑定后,已购买的实例不会释放,但是您可以将其绑定到别的站点,为别的站点提供检测 服务。

• 续费:为目标实例续费,可以延长其使用时长。

### 查看检测结果

参照以下步骤,查看您的站点首页监测和全站检测的结果:

- 1. 登录云盾内容安全控制台。
- 2. 在左侧导航栏,选择站点检测。
- 3. 在首页监测和全站检测页签下,分别查看最近一次检测中发现的风险。

3451773423							
0.00			C ATTAIN				
a mpan	*** 无风险 \$日日始測 2 个站点離页 15 2	¢	- 1995-1997 - <b>无风险</b> - 今日日始期 1 个站街 4 个网页				
前页台	N 2562N						
风险黄型。	25 <sup>5</sup>	× R	2月1日 2月1日	12/1988: 28		× 諸索 URL	12 B
透耀	经市场名	URL		风险结果	风险内容	检测时间	操作
	bb.bqbq01.com	http://bb.bc	(bq01.com/index.php?	新无风险		2018-01-19 01:54:02	风险最否处理地华? 已处理 纠 输
	bb.bqbq01.com	http://bb.bc	(bq01.com	<b>新无风</b> 险		2018-01-19 01:54:02	风险量否处理电路? 已处理 纠 糖
	bb.bqbq01.com	http://bb.bc	bq01.com/index.php?=PHP8885F240-3CR	15元548		2018-01-19 01:54:02	风险是高处理完毕? 巴处理 纠 體
28	902402 <b>60</b>	反清					总具4个结果 < 🚹 >

- 4. 通过提供的 URL,进一步查看并确认风险。
  - 消除风险后,单击已处理,完成处理。
  - 如果您对结果有异议,您可以单击纠错或问题反馈,通过表单将问题反馈给我们。在确认问题后,我们将在算法层面进行优化改进。

### 设置首页防篡改基准和重点监控URL

对绑定站点开启检测时,系统会抓取当前首页,作为判断首页是否被篡改的基准。若您更新过首页 内容,建议您设置首页防篡改基准,重新抓取当前首页。如果您的网站内容很多,您担心在检测中 会将重要的URL遗漏,您可以自定义重点监控URL,系统会优先检测您添加的URL。

参照以下步骤,为已启用站点检测的站点,设置首页防篡改基准和添加重点监控URL:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 站点检测页面。
  - 设置首页防篡改基准
    - 1. 选择目标实例,在其操作选项中,选择设置首页防篡改基准。
    - 确认当前首页基准。如果您想更换首页基准,单击重新获取当前首页,稍后即可查看到系统已重新抓取当前首页,作为首页基准。
  - 添加重点监控URL
    - 1. 选择目标实例,在其操作选项中,选择添加重点监控 URL。
    - 2. 在对话框的输入区域,输入您想要添加的URL,每行一个URL,使用回车换行。最多支持 添加5,000条。
    - 3. 输入完成后,单击提交。

### 自定义词库

在站点检测时,如果您需要对特殊的词汇进行专门识别和防控,您可以自定义词库并将特殊关键词添加进来,进行黑名单防控。具体步骤如下:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 站点检测页面。
- 3. 单击创建词库。

📕 说明:

最多支持创建10个词库。

4. 在创建词库对话框中,输入词库名称,并选择应用该词库的实例,然后单击确定。



在检测站点时,只有您选择的实例才会应用该词库。

- 5. 选择新创建的词库,在其操作选项中,选择管理词库。
- 6. 单击新增关键词,并按照页面提示输入或导入关键词。
- 7. 单击确定,完成添加。

自定义图库

在使用站点检测检查图片时,您可以将特定图片定义为白名单/黑名单图片,进行过滤/防控。具体步骤如下:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 站点检测页面。
- 3. 单击创建图库。



最多支持创建10个图库。

4. 在创建图库对话框中,完成相关配置,然后单击确定。配置说明如下:

配置	说明
图库名称	输入一个用于识别此图库的名称。
使用场景	选择智能监黄或暴恐涉政识别。

配置	说明
图库类型	选择黑名单或白名单。黑名单图库用于特殊防控不良图片,白名单图库会在检测中忽略并过滤您添加的图片。
选择实例	选择应用该图库的实例。
	<b>说明</b> : 在检测站点时,只有您选择的实例才会应用该图库。

5. 选择新创建的图库,在其操作选项中,选择管理图库。

6. 单击选择本地图片,并上传本地图片至当前图库。

### 消息通知设置

云盾内容安全的默认消息推送每天触发1次。您可以设置消息接收方式、账号、和接收时间,也可 以开启/关闭首页风险实时通知。具体步骤如下:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 消息通知页面。
- 设置风险预警的通知接收账户(即接收邮箱地址和手机号码),勾选相应的通知方式(邮件、短 信、站内信),和定期推送时间。

设置			
消息通知	站点检测	OSS 违规检测	内容检测 API
<b>通知接收账户 (</b> 摄	醒邮件和短信将发送到1	以下账户)	
	邮箱地址:		
	手机号码:	<u>_</u>	
通知方式			
	提醒方式: 🔽 邮	件 🔽 短信 🔽 站内信	
通知内容			
内容安全定期推进	送时间(GMT+8): 09:00	~	
	实时消息提醒: 🗌 站;	点 <b>检测首页风险</b> (针对单个	域名风险,每天最多通知一次)
保存			

4. 考虑到首页风险的重要程度,您也可以设定是否开启首页风险实时通知。

开启后,系统一旦检测出首页存在风险,会实时发送消息给您。如果多次检测到风险,为避免您 被打扰,每天最多发送3次提醒。

# 3 OSS违规检测

内容安全通过人工智能技术鉴别OSS中的违规图片,帮助您减少90%以上审核人力,有效降低涉黄 涉政风险。

观看以下视频,快速了解内容安全OSS违规检测功能。

前提条件

- OSS违规检测服务只向阿里云OSS用户提供,在使用前,请确保您已开通阿里云OSS服务。
   如果您还未开通OSS,请前往OSS 控制台,开通OSS服务。
- 您只有通过RAM授权云盾内容安全读取OSS Bucket的权限后,才能使用OSS违规检测。授权步骤如下:
  - 1. 登录云盾内容安全控制台。
  - 2. 在左侧导航栏,单击OSS违规检测。
  - 3. 根据页面提示,单击授权,授权云盾内容安全检测您的OSS图片。



如果您已授权内容安全检测 OSS,则此处没有该提示。

4. 单击同意授权,完成访问授权。



- 如果您开启了OSS 防盗链,您必须在OSS 控制台将https://yundun.console.aliyun.
   com和https://yundun.console.aliyun.com/?p=cts增加至Refer白名单。
- OSS违规检测支持的endpoint包括: oss-cn-hangzhou.aliyuncs.com、oss-cn-shanghai.aliyuncs
   .com、oss-cn-qingdao.aliyuncs.com、oss-cn-beijing.aliyuncs.com、oss-cn-shenzhen.aliyuncs
   .com。

### 设置增量扫描

您可以对指定的 Bucket 中的指定图片(或视频)启用增量内容自动检测,并根据检测结果执行相应处理;也可以对 OSS 空间中的指定图片(或视频)进行存量内容一次性检测,等待扫描完成后,直接查看检测结果,并执行相应处理。

参照以下步骤,对指定 Bucket 设置增量检测计划:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > OSS 违规检测。
- 3. 在Bucket 设置下,从左侧待选择框中勾选需要检测的Bucket,将其添加到右侧的已选择框中。



左侧待选择框中罗列了当前阿里云账号在OSS中的可检测的Bucket,上面的数字表示可选择的Bucket的数量(例如下图中的"15")。

Bucket 设置①					
15	待选择		9	已选择	
	<b>^</b>				
		< .		7	
		$\rightarrow$		7	
				7	
	-			7	
图片检测	视频检	<b>金则</b> 测末:	开启		
保存 ✓ 我E	已同意 《OSS i	<b>违规</b> 检	测服务条款》		

说明:

右侧已选择框中罗列了已选择的Bucket,上面的数字表示已选择的Bucket的数量(例如上图中的"9")。单击已选择的Bucket右侧的漏斗图标可以设置过滤规则,仅扫描指定前缀(或目录下)的图片文件。例如,添加img/test\_,则表示只扫描该Bucket中以img/test\_为前缀的图片。

过滤规则	
扫描指定前缀的图片文件。	
img/test_ × img/201805/test_ × + 添加	
如输入img/test_,则表示仅扫描img/test_前缀的图片文件。如果文件在目录内,文件名前需加上目录路径,整体作为前缀。最多支持 10 个规则。	
确定取消	

如果要扫描的图片文件在特定目录下,您可以在文件名前加上目录路径,以整体作为前缀。 例如,您要扫描的文件在*img*/201805目录下,且前缀为*test\_*,您可以添加*img*/201805/ *test\_*;如果您只想扫描*img*/201805目录下的图片,您可以添加*img*/201805。

	<b>〕</b> 说明:	
	您最多可以为一个Bucket设置10条过滤规则。	
4.	选择开启图片检测、视频检测,并完成相关配置	。参数说明如下:
	图片检测	视频检测 🔵

检测类型 单个服务 元/千张 图片,涉黄与涉政检测分开收费,支持流量包优惠,查看详细价格 ☑ 涉黄 ☑ 涉政	<b>检测类型</b> 单个服务 元/千张视频帧,涉黄与涉政检测分开收费,支持流量包优惠,查看详细价格 注:截帧数=min[下取整(视频时长/截帧频率)+1,最大截帧数] ☑ 涉黄 ☑ 涉政		
自动冻结 ① 当检测分值高于指定分值时目动冻结图片	视频截帧设置 🕕		
开启自动底结	截帧频率 3 秒1帧(频率越高,识别准确率越高)		
沙黄自动冻结阈值 100 分	单视频 是大载帧数 100 帧,超出不再截帧		
涉政目动冻结阈值 99.01 分	仅扫描 100 M大小以内的视频文件		
高级设置 ^			
限定每日图片扫描上限	<b>自动冻结</b> 开启后,违规视频将被自动冻结		
每日扫描量上限 0.2 万	涉黄视频		
	涉政视频		

功能	配置	说明
图片检测	检测类型	勾选涉黄、涉政。
	自动冻结	开启自动冻结后,当检测分值高于指定分值时,自动冻结图片。开启后,需要分别设置涉黄自动冻结阈值和涉政自动冻结阈值。

		<ul> <li>说明:</li> <li>请慎重修改冻结阈值。默认冻结分数是100分,正常 情况下不建议设置到99分以下,分数过低可能会导致 正常图片被冻结,尤其是开启存量扫描时。冻结后的 图片前台不可访问,帮助您防止风险外露。对于被冻 结的照片,您可在OSS违规检测页面将其删除或者解 冻。</li> </ul>
	高级设置	开启/关闭限定每日图片扫描上限,并设置每日扫描量上限。
		<ul> <li>说明:</li> <li>默认没有上限。如果设置了扫描上限,扫描数量超</li> <li>出限制后将会停止扫描,存在巨大违规图片外露的风</li> <li>险。常规情况下,不建议您设置扫描上限。</li> </ul>
视频检测	检测类型	勾选涉黄、涉政。
	视频截帧设置	<ul> <li>• 截帧频率:截帧频率越高,识别准确率越高。最低 60秒1帧,最高1秒1帧。</li> <li>• 单视频最大截帧数:设置超过多少帧以后不再截 帧,作为上限保护参数。最少100帧,最多9999 帧。</li> <li>• 视频文件大小:设置仅扫描多少M以内大小的文件。 检测视频大小不能超出500M。</li> </ul>
	自动冻结	开启后,违规视频将被自动冻结。勾选涉黄视频、涉政 视频。

5. 勾选我已同意《OSS 违规检测服务条款》,并单击保存,保存后的设置即时生效。

设置成功后,系统会按照配置,自动对已选择的Bucket进行图片/视频检测。

### 执行存量扫描

您可以对指定Bucket中的指定图片(或视频)启用增量内容自动检测,并根据检测结果执行相应处理;也可以对OSS空间中的指定图片(或视频)进行存量内容一次性检测,等待扫描完成后,直接查看检测结果,并执行相应处理。

参照以下步骤,一次性扫描指定Bucket中的存量图片(和视频):

1. 登录云盾内容安全控制台。

- 2. 前往OSS 违规检测 > 存量扫描。
- 3. 单击开始扫描,并确认。

?	上次扫描结果将被清空,用于 开始扫描?	展示本次扫描	结果。确定
		取消	确定

4. 在存量扫描设置对话框完成相关设置。配置说明如下:

存量扫描设置						×
Bucket 设置 ①						
8	待选择		0		已选择	
hofuaxin	•					
caffe-bucket		<				
shukuntest	- 1	>		Not Found		
emr-es	- 1					
mtsbucketin	•					
时间范围: 2000-01-	.01 ~	2018-0	)5-11 Ē	1		
文件类型: 图片(1.87	€/千张)	一视频	〔(3.25元/千引	 K )		
athtima 1	秒1帧(频	· 逐越高,	识别准确率越高	5 )		
单视频 100		百共志		- /		
最大截帧数	」 帜,趋田个	丹截帜				
仅扫描 100	M大小以内	的视频文	:(牛			
自动冻结						
违规图片 🦳 🗌						
自动冻结阈值:涉黄 99.01	自动冻结阈值:涉黄 99.01 涉政 99.01					
违规视频						
单个服务1.8 元 / 千张 图片,,3.25 元 / 千张 视频帧 , , 涉黄与涉政检测分开收费 <mark>流量包</mark> 优惠 , 查看 <mark>详细价格</mark> 注 : 截帧数=min[下取整(视频时长/截帧频率)+1, 最大截帧数]				检测分开收费	, 支持	
				取消	开始扫描	

配置	说明
Bucket设 <u>置</u>	选择要扫描的Bucket,并为已选择的Bucket设置扫描过滤规则。具体操作见设置增量扫描中步骤3说明。
时间范围	选择要检测的时间范围,该时间指图片上传到Bucket中的时间。
文件类型	勾选要检测的文件类型:图片、视频。
检测类型	勾选要检测的场景类型:涉黄、涉政。

配置	说明
	<ul> <li>说明:</li> <li>涉政和涉黄检测分开计费。例如,一张图片选择涉政和涉黄检测,既</li> <li>要计入涉政的计费,也会计入涉黄的计费。</li> </ul>
其他设置	如视频截帧、自动冻结等,参考设置增量扫描中步骤4说明。

5. 单击开始扫描,并等待扫描完成,扫描结束后您会收到短信通知。

在存量扫描过程中,您随时可以终止扫描。手动终止扫描后,扫描任务及配置会为您保 留7天,供您在合适时间继续扫描。

OSS 违规检测 点击查看使用说明,如	果您开启了 OSS 防盗链,请至 <b>OSS 管理控制台</b> 将 https://yundun.console.aliyun.com
增量扫描	存里扫描
× 终止扫描	7 存量扫描进行中,扫描结束会短信通知。已扫描约0张图片,0个视频

### 检测结果处理

检测结束后,您可以在控制台查看、或按指定条件查询检测结果,并根据检测结果执行对应操作。

📋 说明:

存量和增量检测结果页面类似,以下以增量检测页面为例进行说明。

- 1. 登录内容安全控制台。
- 2. 前往OSS违规检测 > 增量扫描页面,直接查看检测结果。



3. 设置查询条件,查询指定结果。

您可以指定图片/视频、涉黄/涉政、分值范围、检测结果、Bucket、图片上传的时间范围来过滤结果数据,直接查看您希望看到的结果(图中标识①)。其中,

- 分值范围:每一张图片下方会有违规检测服务检测出的风险分值,分值从0-100,分值越高,代表该图片是违规概率越大。
- 检测结果:违规、疑似、正常。
- 4. 将鼠标放置在每一张图片/视频上,会显示与该图片/视频对应的操作项。您可以进行以下操作:
  - 单击违规并删除,可将图片/视频从内容安全控制台和OSS Bucket中一并删除(图中标识
     ②)。
  - 单击正常并忽略,则忽略该检测结果。忽略后该图片/视频将不再在控制台展示,并不影响保存在OSS Bucket中的图片/视频(图中标识②)。
  - 单击图片/视频可将其选中,并支持多选。单击全选可以选中当前页的所有图片/视频,单击删除所选可将选中的图片/视频全部删除;单击忽略所选可将选中的图片/视频全部忽略(图中标识③)。
  - 若您设置了自动冻结功能,则还可以将已冻结的图片/视频解冻(图中标识③)。
  - 将鼠标放置在感叹号(!)上方可查看图片/视频的基本信息,包括创建时间、文件名和所在
     Bucket(图中标识④)。

# 4 内容检测API

### 4.1 授权访问MTS服务

在提交视频内容检测任务时,如果您选择通过OSS地址(oss://xxxx)上传视频URL的方式,则云 盾内容安全对上传的OSS视频自动截帧。内容安全调用阿里云媒体处理服务(MTS)进行视频截 帧,避免公网访问用户数据,最大限度降低流量费用。您必须授权MTS服务以内容安全的身份递 交视频截帧任务。该操作通过阿里云访问控制中的角色管理功能实现,本文介绍了您需要完成的步 骤。

### 背景信息

您只能在提交视频异步检测任务时,选择上传视频URL的方式进行配置。

关于访问控制服务的角色功能,请参考角色。

通过完成操作步骤,您将实现以下目的:

- 在您的阿里云账号下创建MTS服务角色,并指定由内容安全的阿里云账号扮演使用该角色。
- 授权所创建的MTS服务角色只读访问您的OSS空间。
- 在通过OSS地址上传视频URL时,按照格式要求拼接生成URL并上传。

这样,内容安全的阿里云账号将扮演所创建的MTS服务角色,调用自身MTS服务,访问您的OSS空间,获取视频内容并对其截帧。

### 操作步骤

- **1.** 创建RAM角色。
  - a) 登录RAM控制台。
  - b) 前往角色管理页面,并单击页面右上角的新建角色。
  - c) 选择角色类型为服务角色。

1:选择角色类型 2:填写类型信息 3:配置角色基本信息 4:创建成功 用户角色 受信云账号下的子用户可以通过扮演该角色来访问您的云资源,受信云账号可以是当前云帐号,也可以是其他云账号。 服务角色	]建角色	×
用户角色 受信云账号下的子用户可以通过扮演该角色来访问您的云资源,受信云账号可以是当前云帐号,也可以是其他云账号。 服务角色 受信云服务可以通过扮演该角色来访问您的云资源。	1:选择角色类型 2:填写类型信息 2:配置角色基本信息 4:创建成功	
<b>服务角色</b> 受信云服务可以通过扮演该角色来访问您的云资源。	用户角色 受信云账号下的子用户可以通过扮演该角色来访问您的云资源,受信云账号可以是当前云帐号,也可以是其他云账号。	
	<b>服务角色</b> 受信云服务可以通过扮演该角色来访问您的云资源。	

d) 选择受信服务为MTS多媒体转码服务。

1:361种包括2022年213月5天型1187311100日日从中4087411874118741187411874118741187411874	1985种巴来型  217和与英型相思  31812时已是小小山也  41838330 组服务,受倡服务将可以使用此角色来访问您的云资源。
皆择受信服务,受信服务将可以使用此角色来访问您的云资源。	個服务,受倡服务将可以使用此角色来访问您的云资源。
图释受信服务,受信服务将可以使用此角色来访问您的云资源。	信服务,受信服务将可以使用此角色来访问您的云资源。

e) 在配置角色基本信息页面,填写角色名称,并单击创建。

Ê	说明:	
该操作	可能需要通过手机验	佥证。
创建角色		×
1:选择角色类型 *角色名称:	2:或如果想信用         3:品质角色度本信用         4:0回初初           mtstoA             采取为140个学习、公司英文学品、数字、或"************************************	
盤注:	2890mtsCLA850時行は同cos	

f) 创建成功后,回到角色管理页面,选择新创建的角色,单击管理。

±−≠ sat

- 9) 在角色详情页面,单击编辑基本信息。
- h) 修改策略内容,将"Service"下的内容修改为"1184847062244573@mts.aliyuncs.com",并单击修改角色。

编辑角色		$\times$
角色名称: 策略内容:	<pre>content-moderation  1 {</pre>	
	修改角色	取消

该操作指定由内容安全的阿里云账号(UID:1184847062244573)扮演所创建的服务角 色,调用其MTS服务。

- 2. 为服务角色授权。
  - a) 在角色管理页面,选择新创建的角色,单击授权。
  - b) 在可选授权策略名称下搜索授权策略AliyunOSSReadOnlyAccess,并将其添加到已选授权策略名称中。

编辑角色授权策略				>
添加授权策略后,该角色即具有该条策略	的权限,同一条	授权策略不能被	重复添加。	
搜索授权 精确授权				
可选授权策略名称	类型		已选授权策略名称	类型
AliyunOSSReadOnlyAccess	۹		AliyunOSSReadOnlyAccess 只读访问对象存储服务(OSS)的权限	
		>		
		<		

该操作授权服务角色以只读权限访问您的阿里云账号下的OSS内容。

- C) 单击确定。
- **3.** 复制角色ARN (Aliyun Resource Name, 阿里云全局资源名称)。

- a) 在角色管理页面,选择新创建的角色,单击管理。
- b) 在角色详情页面, 查看并复制其Arn。

基本信息	
角色名称 mts-to-A	信注 接权mtsQLA的最份边间oss
0180101 2017-08-02 15:56:52	Am acsorante:

4. 对要检测的OSS视频对象,按照以下格式拼接生成视

频URL:oss://arn@bucket.region/object

例如,假设您在深圳OSS的bucket foo上有视频对象video/bar.mp4需要检测,则拼接生成的URL为oss://acs:ram::xxxxxxxxxxx:role/mts-to-a@foo.cn-shenzhen/video/bar.mp4(xxxxxxxxxxx是您的16位阿里云ID。)



目前支持的区域(region)包括: cn-hangzhou、cn-shanghai、cn-beijing、cn-shenzhen。

5. 提交视频检测任务时,上传拼接生成的URL作为检测对象。

### 4.2 数据统计

您可以在内容安全控制台查看内容检测API的调用数据统计。

背景信息

内容安全控制台汇总了内容检测API的调用统计数据,支持查询最近30天内调用图片、文本、视频 检测接口的总次数,以及不同检测场景的结果中确认违规量、疑似违规量、和正常量。

### 操作步骤

- 1. 登录内容安全控制台。
- 2. 前往内容检测API > 数据统计页面。
- 在数据统计页面,选择查询时间,并单击查询。支持查询的时间段为最近30内。您可以直接选择最近7天或最近30天;也可以分别选择一个起始日期和结束日期,查询在此时间段内的调用统计数据。

2018	2018-07-26 ~ 2018-08-09													
« ‹		20	18年	7月						20	18年:	8月		> »
_	Ξ	Ξ	四	五	六	日		_	Ξ	Ξ	四	五	六	Ξ
25	26	27	28	29	30	1		30	31	1	2	3	4	5
2	3	4	5	6	7	8		б	7	8	9	10	11	12
9	10	11	12	13	14	15		13	14	15	16	17	18	19
16	17	18	19	20	21	22		20	21	22	23	24	25	26
23	24	25	26	27	28	29		27	28	29	30	31	1	2
30	31	1	2	3	4	5		3	4	5	6	7	8	9
最近7天 最近30天														
											Ë	译时间	目び	定

4. 选择要查询的检测接口类型:图片、文本、视频,分检测场景查看调用统计图表。

数据统计	自助审核							
2018-08-16	~ 2018-08-22	■ Q 查询					图片 爻	(本 <b>视频</b>
视频鉴黄调用统计	+					7 王台洞田目		
1						/ 大志响用国	<u>e</u> :U	
8								
调用量								
					● 确认违规量			
0					<ul> <li>疑似违规量</li> <li>正常量</li> </ul>			
0	08-16 08-17	08-18 08-19 确认违规量 • 疑似违规量	08-20 08- ● 正常量	21 08-22				

### 4.3 数据回流

为了持续改进检测效果,您可以通过数据标记回流,帮助内容安全API的检测模型针对性地学习您的审核标准,提升识别准确度。您可以使用阿里云自助审核平台和您自己的审核平台。

### 使用阿里云自助审核平台

如果您没有自己的审核平台,建议您使用我们的自助审核平台进行审核和数据回流,审核结果将通过回调接口通知您。

### 标记样本

通过内容安全自助审核平台,您可以实时查看图片、文本和视频的识别结果,并对审核中发现的识别有误的样本进行标记,反馈给我们。如果您的审核人力有限,建议您重点审核被识别为违规(block)和疑似(review)的样本。

参照以下步骤,查看并标记识别结果:

- 1. 登录云盾内容安全控制台。
- 2. 前往内容检测API > 自助审核页面。
- 3. 按照以下方式,进行标记:

云盾 • 内容安全	
🔲 站点检测 🏧	*************************************
☑ OSS 违规检测	数据统计
🗋 内容检测 API	图片 文本 视频
◎ 设置	识J场晨:       全部       >       审核结果:       全部       >
	时间范围: 2018-04-19 00:00:00 ~ 2018-04-25 23:59:59 □TaskID: DataID:
	の意思
E	
	最初: pass 最初: block 思想: block
	正常 违规 正常 违规 正常 违规
	특징: block 용장: block 용장: block 용장: pass
	■ 全选 挑量请空标记 挑量标记正常 批量标记违规 提交 <sup>4</sup> 总共 77284 个结果 < 1 2 3 4 5 ··· 3885 >

- 对于您认为正常,却被识别为违规(block)或者疑似(review)的样本,标记为正常(上图标识①)。
- 对于您认为需要管控,却被识别为疑似(review)或者正常(pass)的样本,标记为违规,并选择违规原因:涉黄、暴恐、广告(上图标识②)。

📃 说明:

支持勾选多张图片,进行批量处理,如批量清空标记、批量标记正常、批量标记违规(上图标 识③)。

 标识完成后,单击提交(上图标识④)。被标记样本以及类似样本的检测结果将会按照您的标记 结果实时纠正,同时会通过下文中的回调接口通知您。

结果通知

在对内容检测API检测结果进行审核时,您可以启动回调动作,并设置一个HTTP(s)接口。这样设置 后,当您提交审核时,我们会通过该接口,将审核结果及系统检测的原始内容推送给您。

参照以下步骤,设置自主审核结果回调接口:

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 内容检测API页面。
- 在自主审核结果通知下,设置回调地址(即回调链接callback)。当您添加完回调链接时,系统 会自动生成一个回调种子(seed)。

📃 说明:

最多只能添加一个回调链接(callback)。

云盾 • 内容安全	设置					
💷 站点检测 🏴	消息通知	站点检测	OSS 违规	し 检测	内容检测 API	
四 OSS 违规检测	自助审核结果通	知				
C 内容检测 API	通过指定接口接收	 自助审核结果详细使用说明				
<ul> <li>② 设置</li> </ul>	回调地址:			<u> </u> 修改	回删除	
	Seed :			自复制	心重置	

关于回调通知参数(callback、seed),参见以下说明。

### 结果回调通知参数(callback、seed)

回调链接(callback)需支持POST方法,传输数据编码采用utf-8,并且支持表单参数checksum和 content。系统将按以下描述的生成规则和格式设置checksum和content的值,调用您的callback接口,返回检测内容。

您的服务端接收到我们推送的结果后,返回的HTTP状态码为200时,表示推送成功,其他的HTTP 状态码均视为您接收失败,我们将最多重复推送16次。

名称	类型	描述
checksum	String	由用户uid + seed + content拼成字符串,通 讨SHA256筧法生成。用户UID即账号ID 您可在阿里

### 回调结果参数的生成规则

		云 <mark>控制</mark> 台上查询。为防篡改,您可以在获取到推送结果 时,按此算法生成字符串,与checksum做一次校验。
content	String	JSON字符串格式,请自行解析反转成JSON对象。 content结果格式参见下文。

### content结果格式

content包含以下两部分内容:

- 扫描结果(scanResult): API调用返回结果中字段及描述。
- 审核结果(auditResult):
  - suggestion:系统审核结果,取值:
    - pass:表示审核为正常。
    - block:表示审核为违规。
  - labels: 审核为违规时的具体原因。

以下是一个content结果示例:

```
{
    "scanResult": {
        "code": 200,
        "msg": "OK",
        "taskId": "fdd25f95-4892-4d6b-aca9-7939bc6e9baa-1486198766695
",
        "url": "http://l.jpg",
        "results": [
            {
                "rate": 100,
                 "scene": "porn",
                "suggestion": "block",
                "label": "porn"
            }
        1
   },
    "auditReult": {
        "suggestion": "block",
        "labels": [
            "porn",
            "ad",
            "terrorism"
        ]
```

}

### 使用您自己的审核平台

如果您有自己的审核平台,您可以直接对接反馈接口,将审核后认为识别有误的样本回流给我们。 收到您的反馈后,我们会在下个版本的模型迭代中将您的反馈数据加入训练。关于反馈接口参数,参见具体接口文档。

模型训练需要积累足够的样本,可能无法立即生效。如果有需要,您可以开启自动加入自定义图库 功能,实时纠正结果。

开启自动加入自定义图库的操作步骤如下:

- 通过工单或者您的商务经理,联系内容安全运营人员,帮助您打开实时回流自定义图库的开关,选择要回流的场景。系统会帮助您在自定义图库中自动创建该场景的回流图库,分为黑名单和白名单。
- 在反馈接口的label字段中,对于您认为正常的样本传normal,将该样本加入白名单;对于您 认为违规的样本传任意字段(建议您使用porn、ad、terrorism等风险字段),将该样本加入黑名 单。

■ 说明:

在内容安全控制台,您也可以对回流图库像对其他自定义图库一样进行管理,只不过不能创建与删 除回流图库。关于自定义图库的操作方法,请参考自定义图库。

### 4.4 自定义图库

在您出现突发性的管控需求,而模型更新时间较长,暂时无法满足需求时,内容安全提供了自定义 图库功能,帮助您达到紧急止血的目的。

### 背景信息

自定义图库分为黑名单和白名单两种:

- 加入黑名单的样本以及类似样本,后续算法返回的suggestion都将为block。
- 加入白名单的样本以及类似样本,后续算法返回的suggestion都将为pass。

### 操作步骤

1. 登录云盾内容安全控制台。

2. 前往设置 > 内容检测API页面,并定位到自定义图库功能区。

云盾 ● 内容安全	自定义图库								
<ul> <li>站点检测 999</li> <li>OSS 违规检测</li> </ul>	用户在使用鉴	用户在使用鉴置接口、暴恐接口、广告接口等图像服务时,可添加自定义图片进行防控,添加的图片会在15分钟内生效 + 创建图库 (共可创建10个,已创建6个,自动回流库不计入)							
C 内容检测 API	ID	图库名称	图片数	使用场景	类型	最近修改时间	biztype	操作	
<ul> <li>② 设置</li> </ul>	808001089		30	智能鉴黄	白名单	2018-04-18 16:50:19	12	管理图库(修改)删除	

3. 单击创建图库,并在创建图库对话框中完成以下操作:

创建图库		×
图库名称	test	
使用场景	智能鉴黄 >	
图库类型	黑名单 ~	
biztype (1)	高级功能,选填	
		取消 确定

- a) 按照您的业务命名图库。
- b)选择使用场景:智能监黄、暴恐暴政识别、广告识别。

### 📃 说明:

该场景参数对应于API调用时通过scenes传入的参数。例如,假如该图库适用于鉴黄场 景,则使用场景选择智能鉴黄;那么,在您调用图片或者视频鉴黄服务时,都会默认启用该 场景的黑白名单。

- c) 选择图库类型:黑名单、白名单、疑似名单。
- d) (可选) 输入biztype。biztype属于高级功能,请根据需要进行设置。如果设置biztype,则使 用biztype调用API时,带有该biztype的自定义图库才会生效。
- e) 单击确定,完成创建。
- 4. 管理图库。



- a)回到图库列表中,选择目标图库,单击其操作列下的管理图库。
- b) 在管理图库页面,您可以维护图库内的图片:
  - 单击选择本地图片,上传图片到图库。

📃 说明:

支持上传png、jpg、jpeg、bmp格式的图片文件,且单个文件需要小于5M,一次最多上 传 20 张图片。

- 勾选不需要的图片,单击批量删除,删除图片。
- 删除与修改图库。回到图库列表中,选择对应图库,单击其操作列下的删除或修改可以分别删除 目标图库和修改目标图库的配置。

### 4.5 自定义词库

在您出现突发性的管控需求,而模型更新时间较长,暂时无法满足需求时,内容安全提供了自定义 词库功能,帮助您达到紧急止血的目的。自定义图库支持将指定关键词加入到黑名单,加入黑名单 的样本以及类似样本,后续算法返回的suggestion都将为block。

背景信息

### 操作步骤

- 1. 登录云盾内容安全控制台。
- 2. 前往设置 > 内容检测API页面,定位到自定义词库功能区。

云盾 ● 内容安全	自定义词库 用户在使用文2	自定义词库 用户在使用文本反垃圾接口、文本关键间接口服务时,可添加自定义关键词进行防控,添加的关键词会在 15 分钟内生效,关键词只支持 utf-8 格式								
🔲 站点检测 🚾	+ 创建词属	+ 创建词库 (共可创建10个,已创建9个)								
I OSS 违规检测	ID	词库名称	场景	关键词数	最近修改时间	biztype	操作			
🗅 内容检测 API	750002	-10-10	文本反垃圾	11	2018-04-25 16:29:04	无	管理词库   修改   删除			
<ul> <li>◎ 设置</li> </ul>	761001		图片广告	0	2018-04-19 15:48:17	无	管理词库   修 改   删 除			

3. 单击创建词库,在创建词库对话框中完成以下操作:

创建词库		×
场景	● 文本反垃圾 ○ 语音反垃圾 ○ 图片广告	
词库名称	最长不能超过20个字	
biztype (i)	高级功能,选填	
		取消确定

a)选择使用场景:文本反垃圾、语音反垃圾、图片广告。



- c) (可选) 输入biztype。biztype属于高级功能,请根据需要进行设置。如果设置biztype,则使 用biztype调用API时,带有该biztype的自定义图库才会生效。
- d) 单击确定, 完成创建。
- 4. 管理词库。

管理词库 5 返回上						
词库名称:广告广告						
+ 新増关線	<b>波</b> 搜索关键词	搜索				
选择	关键词	历史命中次数	创建时间	操作		
	法~論	0	2018-04-25 16:29:04	册 除		
	法~倫	0	2018-04-25 16:28:59	删除		
全选     批量删除     总共11个结果     ✓     1						

- a)回到词库列表中,选择目标词库,单击其操作列下的管理词库。
- b) 在管理词库页面, 您可以维护词库内的关键词:
  - 单击新增关键词,按照页面提示在图库中增加要阻截的关键词。
  - 勾选不需要的关键词,单击批量删除,删除关键词;也可以单击不需要的关键词下的删
     除,单独将其删除。
- c) 已添加的关键词,您可以查看其历史命中次数,也可以使用搜索进行定位。
- 删除与修改词库。回到词库列表中,选择对应词库,单击其操作列下的删除和修改可以分别删除 目标词库和修改目标词库的配置。