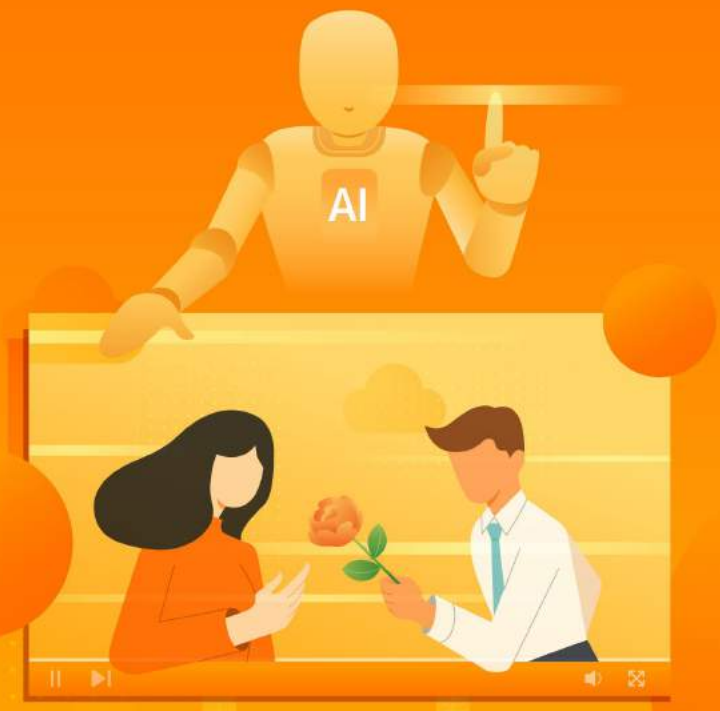
阿里云开发者社区 | 阿里文娱 技术

让刷剧更带感
文娱背后的技术较量

阿里巴巴文娱前沿技术精解



AI弹幕、VR技术、5G超清，
8位阿里技术专家详解智慧文娱关键技术

藏经阁

阿里云开发者电子书系列

加入交流群

阿里巴巴文娱技术
钉钉交流群

关注我们

阿里巴巴文娱技术公众号

目录

5G 篇

概览　阿里文娱面向 5G 的技术布局

5G 和 AI 是业界火热的话题，行业普遍认为这两项技术可能成为下一次工业革命的开端。

5G 作为新一代的通信行业标准，理论上的传输速度可达到 Gbps 量级，去年正式开始商业化。5G 的传输速度可达百兆量级以上，低延迟以及大带宽的特性将对视频内容的形态和交互方式带来变革。

而机器学习作为技术界和学术界都炙手可热的领域，无论是搜索推荐还是视频内容理解，亦或是人脸识别或自动驾驶，都是 AI 大展拳脚的领域，这些技术也逐渐服务我们生活的方方面面。

我们认为，5G+AI 会推动视频内容的生产和消费产生新趋势，主要体现在视觉体验、内容形态和交互方式这三方面：

第一，在视觉体验上，超清高清视频的普及度会加大，像 4K、60 帧或者 HDR 10bit 级的视频内容载体，将为用户提供更好的视觉体验；

第二，在内容形态上，视频形态创新的前沿领域，典型的如 AR、VR 技术以及优酷一直在深入探索的 6DoF 技术，依然要求近百兆带宽才能达到极致的体验水平。而现在，随着 5G 的发展，这些新的内容形态将更好落地。

第三，在交互形式上，5G 的技术特性同样孕育出新的交互形态。例如 3D 视频、视频游戏化、云游戏等。更低延迟带来更实时的交互，更大的传输速率能够承载更多的视频信息，也为 6DoF、子弹时间这种新的交互形态带来落地的可能。

阿里文娱高级技术专家 | 肖文良

第一章 5G 超高清的三大关键技术

作者 | 阿里文娱高级算法专家 张行

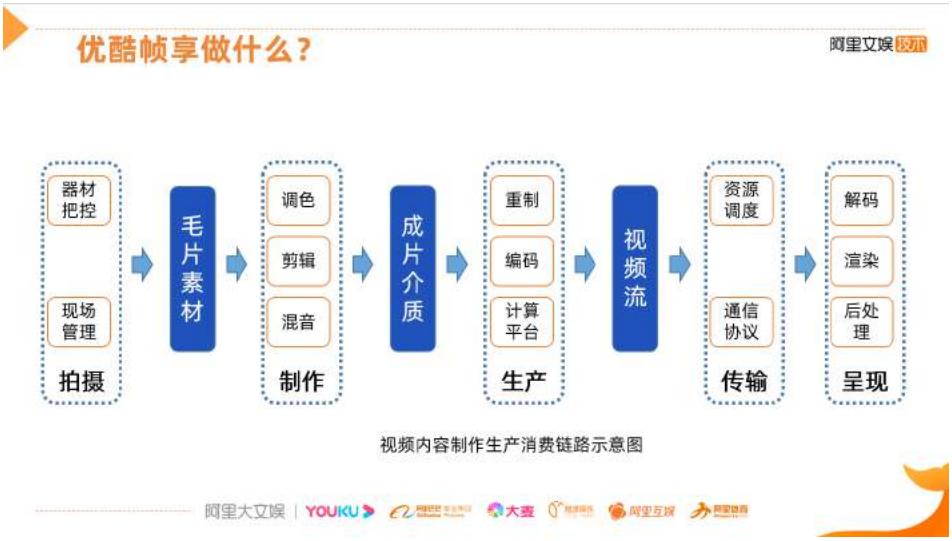
随着 5G 落地，用户对视频体验的要求越来越高。在带宽不是超高清的主要矛盾之后，超高清还存在哪些挑战？我们距离全面超高清还有多远？阿里文娱一直在相关的技术预研，在 2019 年底，推出了互联网视频行业超高清解决方案——帧享。那么，帧享是什么、有哪些关键技术、未来有哪些发展方向？

一、什么是帧享？

帧享是一个超高清的解决方案，从 2B 到 2C 的视角，帧享有 4 个技术能力：

- 一是高帧率增强，可提供最高 120 帧的超高帧率视频，顺滑的呈现物体运动场景。
- 二是超高分辨率，对于画面中微小的细节与结构，在帧享的视频中也能刻画的非常清楚。
- 三是 HDR 高动态渲染，画面对比更丰富，颜色鲜活有质感。
- 四是帧享环绕音效，我们利用声道间的相位差异，充分体现声音的立体感和空间感。

前三个方向的特性分别体现了帧享对于时间、空间、亮度、色度的超高分辨与呈现能力；第四点是声音特性，声场效果，这四点组合起来，既是帧享能给用户提供的关键体验特性，也涵盖了观众对于超高清的诉求。



要真正将帧享落地，需要深入到视频制播产业的各个环节中，从左到右有 5 个关键词：拍摄、制作、生产、传输和呈现，这 5 个环节环环相扣，每一步都决定了最终视频的呈现质量。我们首先要保证每一步都能够正确地处理，尽可能采集和保留更多内容信息；其次是挖掘链路上各环节的处理能力，利用我们在制作、生产和呈现上的人力和算力，进行信息的重建和增强，提升视频体验。

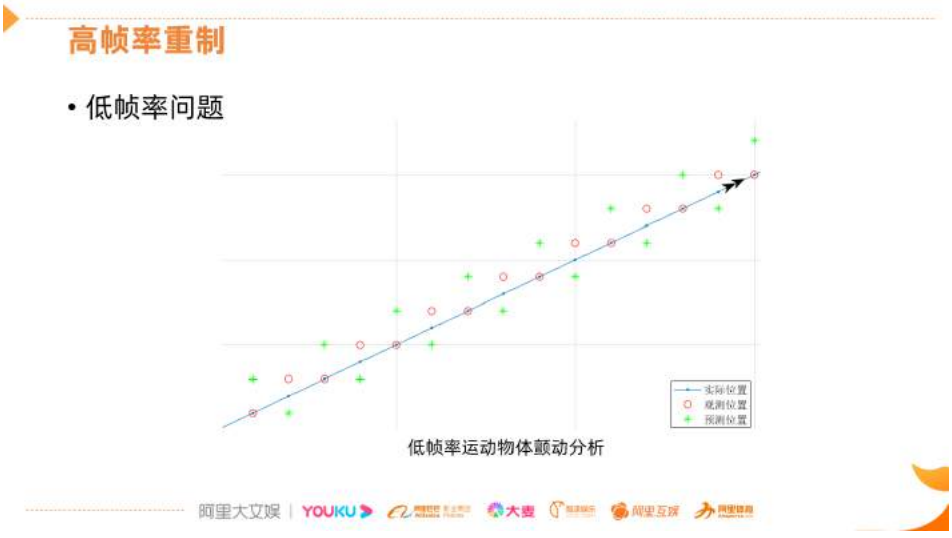
具体来讲，在拍摄和制作环节，我们会给出明确的超高清视频的要求规范；在制作环节，开放云剪辑能力，为后期的剪辑提质提效；在介质环节，做严格品控，保证介质内容的基础质量。在生产环节，减少转码的损失，利用我们平台的计算能力进行恢复和重置增强，同时对视频进行结构化分析，拿到视频的各种分类、场景、标签等高低层的语义信息，将其与码流一起传输到终端设备上，并进行适配的后处理增强和渲染。这种适配包括对内容、设备和用户偏好的适配等，确保最终的体验和效果。

二、帧享的关键技术：高帧率重置、高动态渲染、云加端增强

1. 高帧率重制

（点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

从视频中可以明显看出，低帧率的竖线运动时一直在颤动，而高帧率的运动就很平滑。为什么低帧率会抖动?



如上图，x 轴表示时间，y 轴表示位移，物体的匀速运动在坐标系中是一条斜线，如图中有箭头标记的蓝线。而实际的物体位置在这条蓝线之上。由于低帧率的刷新率是有限的，物体的实际位置在一帧内是固定的，到下一帧会跳跃到另一个位置，就像上台阶一样。人的眼睛会天然的跟踪运动的物体，也会根据当前位置和运动速度，去推测物体的下一个位置，如图中绿星星所标记的。我们看到物体的实际位置和物体的预测位置一直不重合，且预测位置一直在实际位置的上下抖动，非常伤害观看体验。

高帧率重置，在算法上就是插帧。插帧技术已经存在很久了，方法大概分成两类，一类是基于特征的传统方法；另一类是基于数据的网络方法。两者思路是一致的，根据像素的帧间相关性去推算光流，再做插值。

在传统算法中，先根据多帧的视频图像去做光流，预测出前后向光流，来映射到需要插帧的相位上。这时候就需要考虑很多特征，比如到底是用前向光流还是后向光流、用双向光流还是单向光流，哪些地方是露出遮挡区域等，根据这些去做插值重建，得到高帧率视频，这是一种完全基于运动特性的传统方法。

网络方法非常类似，只是将光流的预测还有像素的差值都用网络来实现，还有一些网络方法可能更极端，它会把光流网络和插值网络合二为一，直接用一个端到端的数据训练，得到一个插帧网络。但无论是传统还是网络办法，在插帧中有一个难以解决的问题——在一些运动的交界处，光流很难严格贴合物体的实际边缘，这样会导致各种各样的问题。

优酷是如何优化的?

首先是基于成熟的插值算法，将各点效果做到极致，在实际场景中有效解决问题；其次是拆解问题，尝试把通用的插帧问题，分层分类成不同的垂类，用不同的插帧方法来解决，实现整体最优。

1）场景分类。在时间上做分类，将时间轴上的一个视频按照场景切开，分成了多个场景，把不同场景分成全局运动场景、静止场景、复杂运动场景、片头片尾等。

2）目标的分割。在空间维度将图像分成多个目标区域，例如台标角标的区域、字幕区域、前景背景、露出遮挡的区域。

3）垂类场景的插帧完成后，再经过一些柔性的融合得到最终的插帧结果。

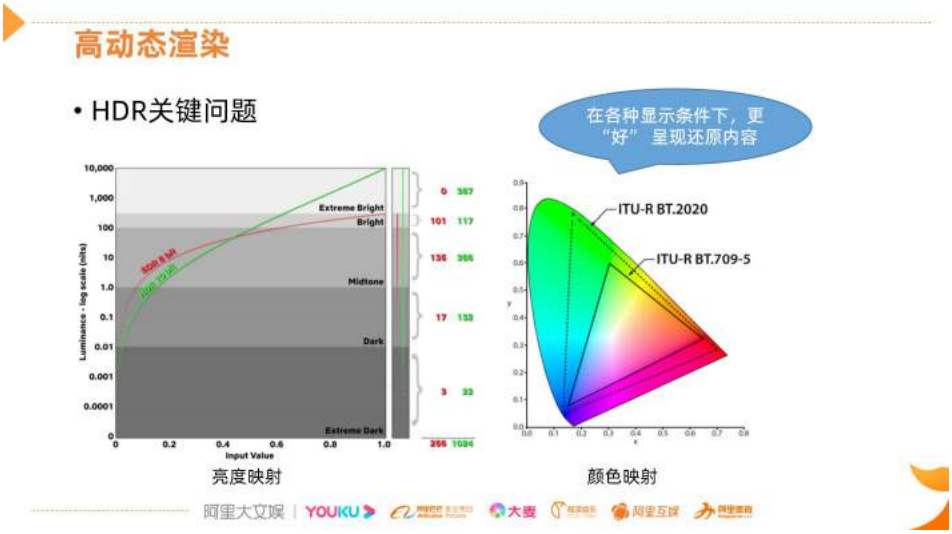
4）人工校对。无论用多么精巧的办法、算法，总会有一些疑难的 case，是技术无法处理的，所以在设计算法时，会自动对疑难 case 进行标记。在审核后台，这些标记区域进行人工审核，对于有问题的插帧结果进行回退处理。

上图是对比图，左侧上方飞掉的字幕，通过对字幕区域的特殊处理以后，已经能够正常做插帧了。右侧，将运动光流进行精细化，让光流更贴合运动的前景轮廓，有效去除在运动物体周报的光圈效应。

2. 高动态的渲染

高动态渲染其实就是 HDR。上图是对比图，左侧是 SDR 效果（画面偏灰，看不清细节）；右侧是 HDR 效果，画面很美，点点繁星和山势的暗部细节轮廓都非常清楚。

HDR 是一个成熟概念，行业中有各种各样的 HDR 标准。我们如何区别中间的差异，并选择一个好的 HDR 算法？ HDR 解决的是一个从高动态到低动态，从宽色域到色域的映射效果问题。自然景物能够呈现出的亮度范围是非常高动态的，从 1/万 nit 到 1 万 nit 以上都有。但显示设备能够呈现的亮度范围是低动态的，大部分只有几百 nit，而低亮也不够低。所以天然的，把自然景物呈现到显示器上，就面临着一个从高动态到低动态的映射问题。所以，HDR 的关键不是 8ibt 还是 10bit，也不是 4k 或者 1080，而是去理解内容和设备，确定在什么设备什么环境下，用什么样的映射去渲染内容，达到主观效果的最优。



上图，左侧是亮度从高到低映射，右侧是色彩映射，需要把马蹄形的大的宽色域映射到内部小三角形上面的窄色域。

帧享 HDR 在技术上做了哪些改进？

一是测屏校屏，帧享要做标准的颜色管理，需要将不同颜色做到在不同设备做到

显示效果一致，用来排除屏幕的颜色偏移，把颜色做的更加准确。

二是屏幕亮度和色度适配，不同设备的亮度差异非常大，从两三百尼特到上千尼特，我们的测试也发现，即使用标准的 HDR 视频，在不同亮度的设备上面的效果也存在差异。 所以帧享 HDR 采用了多种的流策略，对于超过 500 尼特的屏幕，输出标准 HDR 流；对于低亮屏幕，基于亮度去适配调整出独特的 SDR 流。

三是内容适配。每一个场景的内容，很少是满动态或宽动态，有的场景整体很亮，有的场景整体很黑，这时我们可以取巧一点，将内容所在的部分亮度范围做更好的映射，然后在其他亮度范围，将映射做的差一些，这就是根据内容来做动态映射的一个出发点。帧享的 HDR 也是基于这一特性，用动态元数据，根据场景做动态的 tone mapping。

四是做链路的把控，后期、平台以及端上渲染，都可以做这种映射，但不能各自为战，需要信息互通、互相协同，用统一的映射将效果做到最佳。

下图是 HDR 对比图。

第 1 幅是颜色准确性、渲染颜色准确性的对比。右下角是优酷在苹果上的播放效

果显示，其他三张都是同一个安卓手机的不同 APP 的显示效果。因为屏幕本身是有些偏色的，所以可以看到友商两幅图的效果，人脸比较红润，就会红的不太正常。 但是优酷，人的脸色比较正常，更像苹果的颜色显示，所以对比就能说明在我们优酷通过测屏校屏，能够去纠正错误的颜色渲染，然后得到更好的颜色效果。

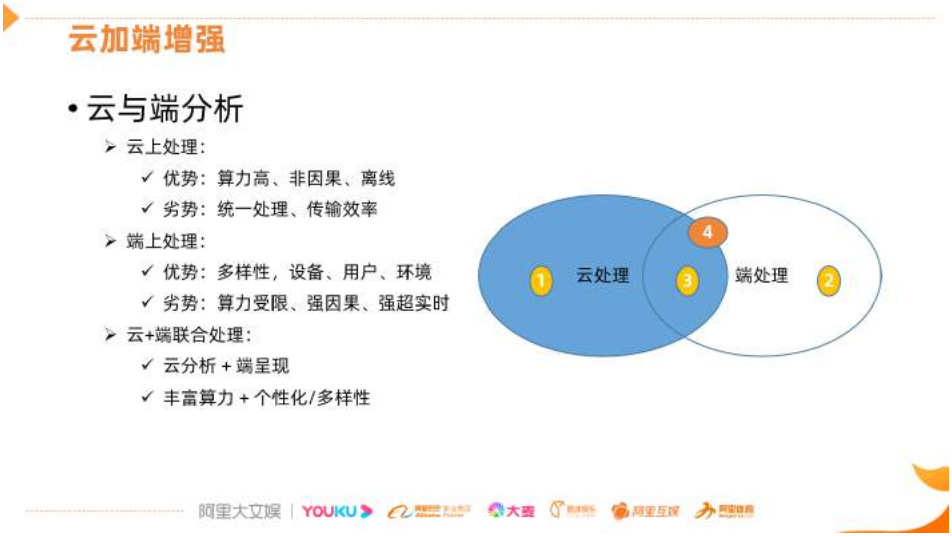
上幅图是帧享 HDR 的对比图，左侧是 HDR 前（画面颜色整体偏亮，对比小、画面偏灰偏白）；右侧是 Tone mapping 后的 HDR 效果，动态 TM 后，扩大对比度，提升了画面质感。

3. 云加端增强

以前，我们常遇到这些问题：为什么视频流很好，到电视上效果不佳？每个设备的效果不一致，如何兼顾？如果知道内容特性，算法参数可以设置的更好，但是无从知晓，所以效果只能打折。以上都反映了一个共同问题，体验是整条链路的体验，必须将云和端协同起来，一起为体验负责。

云和端如何做协同？云上，在编码前做前处理；端上，在解码后做后处理。我们在云上处理的优势，主要是算力丰富、算力高，并且它是非因果和离线的，可以算得

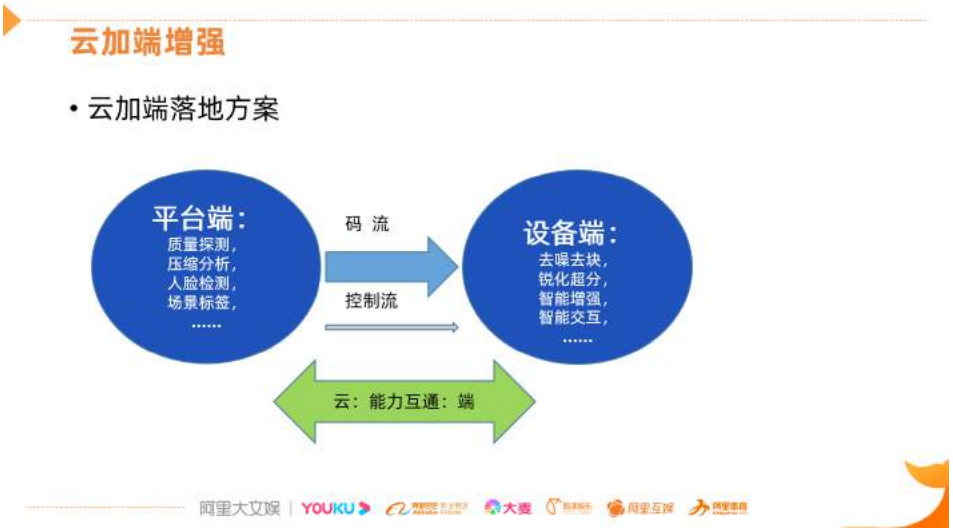
很慢。劣势是云上算的时候，不知道设备信息，所以只能够去做统一的处理，不能单独调优。其次，云上的增强恢复重建，都是增加信息量，所以压缩效率低，压缩后的码率高，导致传输效率降低。在端上，我们知道设备、用户以及环境的信息，用多参数、多种算法做适配，是一个多样性的能力。



我们将云和端联合在一起，用云上的丰富算力做分析，用端上的多样性做呈现，实现优势互补的效果。右图的 4 种情况，1 是纯云端的处理，2 是纯端上的处理，3 是云端都可以处理，4 是云加端的协同处理。

云 + 端的联合处理到底有哪些应用？

基于算力优势，我们会在云端做复杂的探测、分析、分类，打标签、编码，再将码流和探测出的语义信息、一些结果通过控制流去传输到设备端。用来指导端上的后处理模块进行参数的设置、算法的选择，以及适配处理。例如，通过去块、锐化、超分等让端上效果更出色。



案例一，去块。块效应是压缩导致的，在码率不够或者低亮的场景中。统一的去块，有可能会损失信号的有用细节，使图像变得模糊。但如果我们能够做云加端的配合，可以在源头将流上块的强度、类型都探测出来，然后把信息传到端上，用这种信息去控制端上的去噪去块算法的强度，达到既有效去块又能够保护细节的效果。

案例二，智能满屏的效果对比。

优酷有大量的年代剧，往往是 4:3 比例，现在屏幕尺寸是 16:9，甚至是 23:9、22:9。如果直接播放 4:3 视频，画幅会很小。普通平铺是以图像的中心为中心，这样的构图布局经常会丢一些重要画面。优酷智能平铺是利用 CV 的识别分析能力，将眼睛更关注的信息保存下来，让画面的布局更合理。

所以整个应用过程就是在云端，利用分析理解能力，对画面进行自动的分析、提取，将信息与码流一起传到端上，根据信息进行渲染窗口的调整，达到实时的拆切满屏的目的。优势是一个流能够满足各种尺寸屏幕的观看需求。

三、优酷的超高清愿景

帧享的愿景是，在 5G 和 AI 的背景下，为国内的互联网视频超高清路线提供解法和答案，推进视频的超高清体验的升级，让 C 端用户早日进入到超高清的观影时代。另一个愿景是超高清产业共赢。我们需要有超高清的标准去约束视频产业链条的各方，制作生产出符合超高清标准的内容、设备，培养提升用户心智，愿意为体验买单。只有用户愿意买单，平台才愿意为超高清买单，制作公司才会愿意为超高清买单，实现超高清的商业化、规模化，实现用户、制作、平台、终端整个链条上的共赢。

第二章 基于人脸识别的 AI 弹幕

作者 | 阿里文娱高级无线开发专家 少廷

有些弹幕比剧情还精彩，那些脑洞大开、观点鲜明的弹幕，让千万用户参与到“剧情创作”中，所以很多人都喜欢边看剧，边看边发弹幕。你发现了吗，在 AI 算法的加持下，弹幕的呈现形式也花样翻新，优酷的很多剧都上线了基于 AI 人脸识别的跟随弹幕，与剧情的贴合度更高，可玩性更高。这类弹幕是如何实现的呢，有哪些核心技术?

在 GMIC 智慧文娱技术专场上，阿里文娱高级无线开发专家少廷，分享了在优酷播放场景中，如何让互动结合算法的识别能力，实现新的 AI 弹幕形态。同时也介绍了优酷在互动游戏化领域的探索，如何让互动与内容相结合的应用实践。

一、播放中的互动场景

我们先思考下，为什么在播放场景中加入互动环节? 归纳起来有四个价值：一是让用户更好的融入剧情，参与到剧情之中；二是互动是实时的，能够对用户做实时反

馈，体感非常好；三是很多用户在看剧时，精力不是那么集中，我们通过互动去抓住用户的剩余注意力；四是增加社区属性，让用户尽情表达，产生归属感。

上图是优酷重点播放互动的一些场景，包括在画面中弹出提示，对相关情节信息做科普介绍；双流播放形态，通过子母屏为播放场景提供第三人称的视角；在剧中加入答疑、猜剧情，投票中等，当然最核心的形态是弹幕。

因为无论从用户规模还是用户效率看，弹幕都是效率最高的互动形式，是互动最头部的场景。说到 AI 互动弹幕，我们的出发点很简单，希望用户的互动与内容能更紧密结合，让互动的展示形态更年轻和二次元。之前我们在弹幕中做过角色扮演的弹幕，所谓角色扮演弹幕，就是用户可以选择剧中角色，以他的口吻和头像来发弹幕。

这种角色扮演的弹幕，转化率非常高。我们就进一步思考，用户能不能直接在角色上，通过类似漫画配图的方式发弹幕，与内容融合到一起。这就是 AI 互动弹幕的由来。

用户的脑洞和创造力是无穷的，大家总是能想到各种很神奇的弹幕文案来调侃，能让用户有发挥空间与内容产生互动。

二、基于机器视觉的互动弹幕的技术挑战

1. 技术面临的问题：识别放到端侧还是云端?

一是识别剧中人物，去识别人像，这个识别本身有成熟的算法，既可以放到端侧，也可以放到云端，也就是服务端。那识别能力应该放到哪里?

核心的识别能力，如果放到 C 端。识别的功耗和性能开销是很大的。如果针对某一些垂类场景，在比较短的时间内识别，放在 C 端完全没有问题。但如果在长视频中，从头到尾都有一个 C 端的识别引擎在跑，对机器的性能、开销、耗电都难以接受。

二是 C 端识别的精准度，因为算法识别直接输出结果，很难达到产品化的要求。在这一过程中，还要对识别结果进行二次加工和处理，包括平滑处理、降噪。这些处理都需要更多的工作时长，如果都放到 C 端，难以保证实时性。所以我们将整个识别引擎都放到云端，通过云端识别输出数据，经过优化处理后把这些数据打包下发到端侧，实现投放和互动。

2. 算法和工程如何对接?

工程和算法如何对接、工程如何解决算法输出后的各类问题?包括识别精度、数据抖动、视频文件变更之后导致的数据不一致的问题；另外，端侧要解决核心的体验问题。在播放过程中，镜头频繁切换时，这些人像在镜头中变化的问题，手机上的界面的横竖屏的适应；还有气泡样式的弹幕，在不同剧中和内容的氛围会不会产生差异，如何去融合等。

以上是我们的技术链路，有下到上，依次是算法侧、服务端、客户端。在算法层，我们通过模型训练，抽帧对视频的每一帧的画面进行识别，通过人脸检测和跟踪来抓取每一帧中的人脸数据，再传输到服务端进行预加工，包括数据合并和降噪，将人脸数据打包，通过互动投放引擎投放到端侧，端侧来实现核心交互和基础体验。

（1）在算法侧，具体的技术细节如下：

1）视频抽帧与识别

抽帧现在按每秒 25 帧来抽取，也是一个视频的基础帧率，对于高帧率的视频当然可以抽的更多，但这与机器开销和耗时是成正比关系的；另外在算法侧，人脸的识别引擎有几部分，一个是标准的人脸识别，能够识别大多数人脸的场景。其次，优酷还针对明星和人物角色，做了单独优化，我们会抽取一些剧中的明星和角色的数据，作为数据集去训练，提升剧中主角和明星的识别率。

2）人脸跟踪

算法会识别出视频中每一帧的人脸和对应坐标，坐标用来标注人脸对应的位置，每个人脸帧也对应有人脸的特征值，相同特征值能够识别为同一个人。这样通过坐标和特征值，我们就能识别出一个镜头从入场到出场的序列帧里同一个人的人脸运动轨迹。

3）平滑处理，防抖动

我们知道算法是基于单帧对人脸做识别的，人脸的位置和大小的识别结果是有像素级的误差的，同一个人脸哪怕运动很轻微，上一帧和下一帧的识别结果的方框是不会严丝合缝对齐的。这样把连续的每一帧的识别结果像放动画片一样串起来，这种像方框一样的识别结果在播放场景中就会看到明显的抖动。

我们在工程侧，对抖动做了检测和计算，将抖动限定在一个范围之内，让整个人脸的轨迹更平滑。

（点击链接 https://developer.aliyun.com/article/765083，观看视频效果）

视频 2 是算法直接输出的人脸轨迹的识别结果，大家可以看到人脸的识别结果是伴随人物运动在抖动的，既有位置的抖动，也有大小的抖动。视频 3 是经过平滑处理后的结果，人物在整个镜头中走动，我们的识别结果输出是稳定平滑的，抖动效果已经被平滑消除掉。

（2）服务端。服务端对算法输出的人脸的原数据要做一系列的处理：

1）降噪，过滤掉镜头中不重要的人脸的杂音。所谓不重要，是因为镜头中有大量的路人镜头，很多是一闪而过的，这些镜头的人脸对于用户交互来说是不具备太大价值。所以我们要把路人的、一晃而过的镜头中的人脸都过滤掉。

2）防抖。我们会对原数据进行平滑处理。

3）合并，就是合并一组连续帧中相同一个人脸的位置，打包输出。比如一个镜头中，一个人从左边走到右边，我们通过特征值识别出来这是相同一个人的连续镜头，那就把这一个镜头的人脸数据打包。这种包含了相同人脸和对应轨迹的数据，合并到一起之后，就成了一个数据包。然后将用户发的弹幕数据跟人脸的轨迹数据包绑定，投放到端侧，端侧负责展示。

3. 在端侧的技术细节

借助优酷互动投放引擎，将人脸识别结果的轨迹数据包和弹幕一起投放，在端测完成数据解析和展示。端侧有一个轮训引擎来轮训互动投放的数据，当播放进度到某

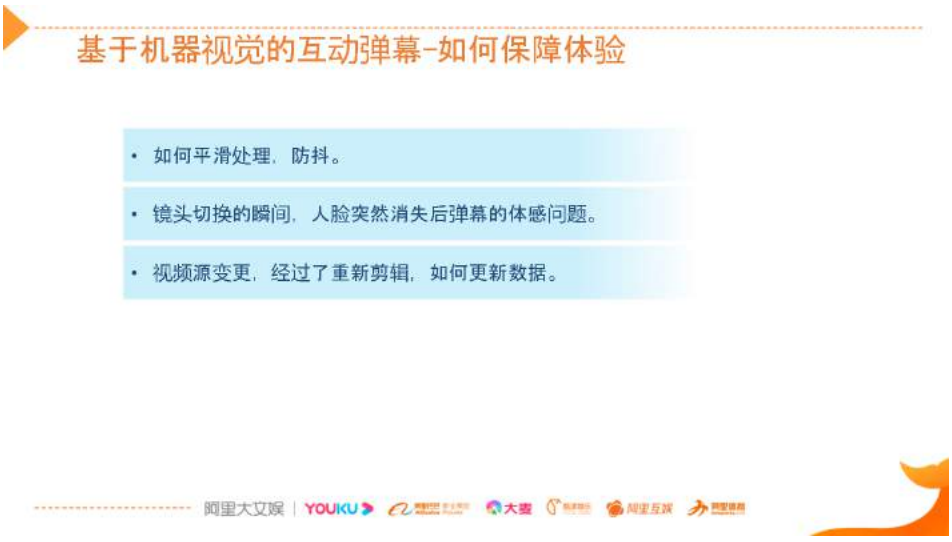
个位置时有人脸轨迹的数据和弹幕数据时，端侧会把人脸位置的弹幕气泡显示到播放场景中。

另外，客户端还要解决技术体验问题，包括镜头切换，虽然是细节问题，但是其实费了很多功夫。 例如，用户想发弹幕，在发送的一瞬间，镜头切换了。在用户侧看到的是弹幕闪一下就消失了，这个体感是非常不好的。

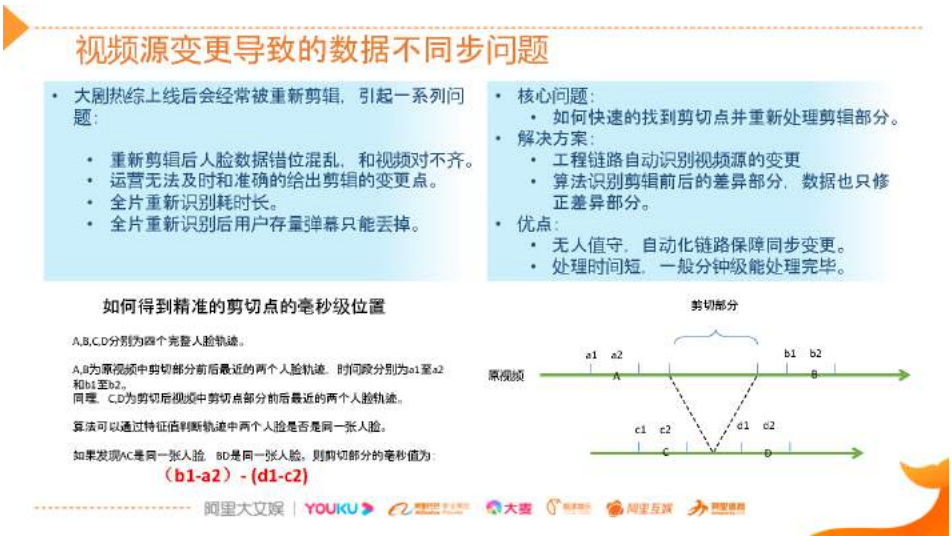
所以，在技术上我们要保证人脸弹幕发送后，会在同一个位置显示一定的时长。

现在是一秒左右；投放弹幕的时候同样如此，如果用户在镜头的最后几百毫秒发送气泡弹幕，我们会把它做时间前置偏移，让弹幕稍微往前，在镜头中完整展示出来。

另外，在氛围方面，我们有古装剧、有现代剧、有悬疑、有综艺，剧的类型的差异比较大。气泡弹幕与播放结合比较紧密，我们还是希望它的效果、样式、氛围能够跟剧本身贴合。我们在端测实现这类样式的动态配置能力，对于不同的剧集能够展示不同的风格。刚才大家在案例中也看到了，有红的、有蓝的、有灰的，然后古装剧会带一些纹理小花之类。



另一个严重问题，当视频剪辑变更之后，数据如何快速更新?



首先为什么会有视频变更？这是视频行业中的普遍现象。例如一个大剧，一个热综，在整个播放周期中，会经常因为一些原因被重新剪辑。重剪后就会引发一系列问题，比如，可能有一个桥段就被剪掉了，原来识别出来的人脸的数据结果和现在新的视频就对不齐。另外，运营同学或者剪辑同学，其实无法准确的告知，在什么时间点剪了什么内容，无法依靠人力来保障同步。

另外，假设我知道视频被剪辑了，全部重新识别一遍行不行？回答是不行，第一是耗时长，额外开销高；第二，如果重新识别，相当于镜头中的人脸的数据全部都更新了一遍，用户发送的存量弹幕和人脸相结合的部分，就无法还原。

所以我们核心问题，就是要解决如何快速找到剪切点，并且只处理剪辑掉的部分。

技术解法：

首先工程侧要自动识别变更，一旦视频源变更，服务端就能接受到相应的通知，能够启动去做重新识别。就是算法要识别剪辑之后差异的部分，找到中间的差异的时间段。这样处理的优势是：1）实现无人值守；2）识别时间短，分钟级处理完毕。

具体是如何实现的？通过算法识别的特征值。例如，原视频中 a 和 b 是两个人

脸，中间的时长是 n 秒，剪辑后变成如下这样；当算法再去识别新视频源时，发现 c 和 d 对应的人脸特征值和 a 和 b 不一致、时长对不上。通过取两者（a 和 b，c 和 d）的差值，就知道那部分是被剪掉了，然后去处理差值的部分，一两分钟重新识别完，保证数据重新上线。

左边是用户用 AI 互动弹幕发布的内容，很多都是优秀的段子手，大家能够玩起来。

右边是弹幕中看到的用户反馈，表达不是很理解、不清楚是做什么的，或者认为弹幕干扰正常观影。所以目前 AI 弹幕还是小范围的投放，并没有全面的铺开，因为我们也在思考用户对弹幕的接受度。 所以在长视频中，并不是所有用户都能接受。更好的方向是将这种互动性更强的更有意思的模式，与短视频、直播结合，因为在短内容和碎片化内容中，用户的接受度会更高，娱乐性会更强；

4. 播放互动的相关尝试和展望

（1）互动交互能够和内容更紧密结合，交互方式更游戏化。

通过对视频内容的识别，能够识别到或者理解到剧中的人物或者剧中的演员，甚至剧中的物体。这样我们对内容本身的场景和角色是能做到理解的。这样就可以将互动投放跟内容做的更紧密，比如说能够跟明星结合的更紧密，甚至能够跟商业化结合的更紧密，让商业化和内容本身结合，这也是一个方向。

（2）互动的效果更炫酷和惊艳。

让交互结果和交互的视觉上更年轻化，更让人有意外，让用户能有更进一步互动

的欲望。

（点击链接 https://developer.aliyun.com/article/765083，观看视频效果）

（3）用户参与的数据沉淀能够反馈指导内容的制作和生产。

大量的用户参与互动的数据能够沉淀下来，通过对数据的理解和二次加工，抽象出一些结论。包括用户在互动上的高潮节点和用户高反馈的点。另外，对内容的理解，希望能够指导内容制作和生产。

（4）短视频、直播这类碎片化的内容能够发挥互动的新价值。

这种现在新的更碎片化的短视频和直播内容中，探索一些新形态和新价值。

第三章 结合 5G 和边缘计算，优酷如何做云渲染?

作者 | 阿里文娱高级技术专家 伊耆

当 5G 来了，视频还是平面的影像吗，只能静静观看吗？一定不是！现在，你可以像玩游戏一样，参与到视频内容当中，还能体验新的播放形式，比如发 AI 弹幕、猜剧情、横竖屏随意旋转，立体的观看进球一瞬间，看到屏幕之外的更大画面等等。这背后的技术是如何实现的，未来有哪些新交互方向?

在 GMIC 智慧文娱技术专场上，阿里文娱高级技术专家伊耆分享了如何利用终端设备的交互特性，结合内容和算法，所实现的新观影模式的探索。同时结合 5G 网络和边缘计算所做的云渲染技术预研。主要分为四部分：

- 一是视频和游戏的共性和差异，如何看待两者?
- 二是视频场景结合内容、算法探索播放新交互模式
- 三是结合 5G、边缘计算和立体视觉的云渲染技术
- 四是未来的思考和总结

一、视频和游戏的共性和差异

为什么要做新交互？其实用户在文娱消费体验上，尤其在视觉体验上，主要集中在两个领域，一是视频，一是游戏。我们在思考切入点时，更多是关注两者之间的共性和差异，寻找结合点。

首先回想一下，你在玩游戏是一种什么样的体验？网上有很多类型的游戏，休闲类、益智类、竞技类等等，我们可以发现游戏的特点是交互性越强，竞技属性越强；交互属性越弱，休闲体验越强。视频的本质相同，在一个纯被动观看的过程中，内容本身会带来感官刺激，但更多体现在休闲上。但随着交互属性的加入，比如当视频引入 VR/AR 等互动后，其形态也更趋于游戏化，更像是一种休闲类的游戏。也就是视

频的“内容属性”与游戏的“交互属性”结合，最终它可能就变成一个像游戏化的视频了，用户会获得比较强的沉浸式的感觉。

二、视频场景的新交互模式

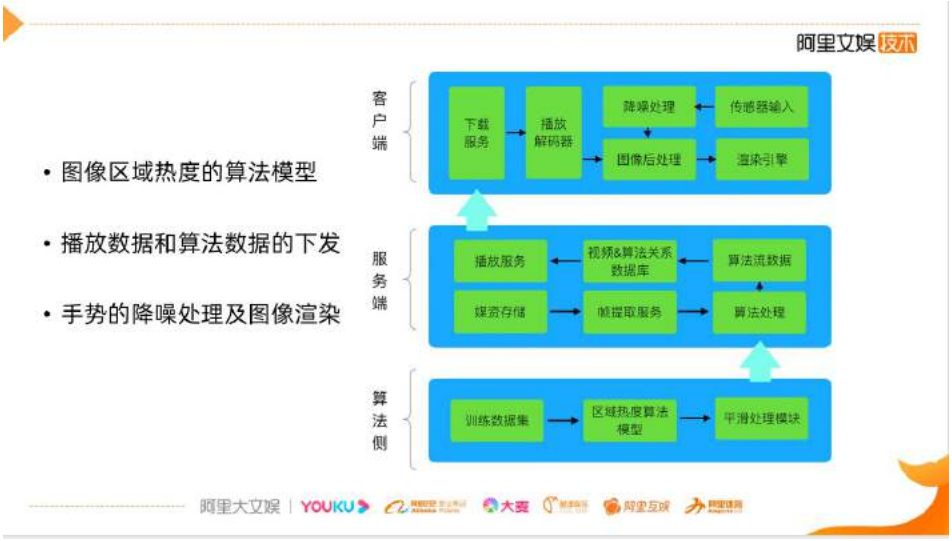
（点击链接 httpsdeveloper.aliyun.comarticle765074，观看视频效果）

参考优酷在互动剧的尝试，我们在播放和交互领域的结合也做了非常多的探索。先看 2 个视频。视频 1 是常规的旋转，体验还算顺滑；视频 2 是加入算法后的策略，在旋转过程中，画面始终是平稳的，甚至用户在横移手机时，可以在屏幕中看到更多的画面，这也是初步尝试。

以旋转的视频为例，形式上看似简单，但它背后也有很多技术点：

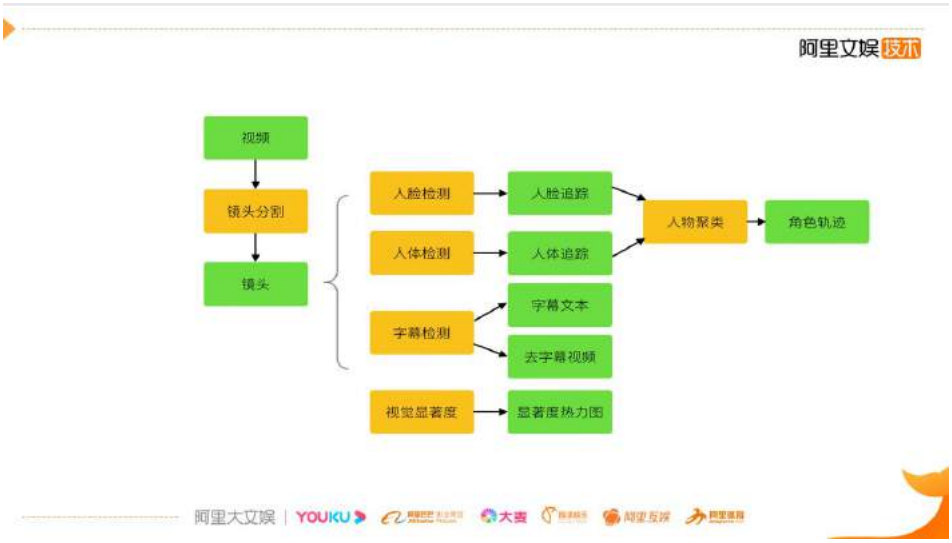
1. 旋转手机时，不丢失画面中心。我们看视频时，视线往往会聚焦在一个中心人物，或是一个场景中心。当手机旋转，自然也不希望丢失视觉的中心点。所以需要一套算法去识别观影中心点。在此基础上，通过服务链路去打通，将算法和视频画面联合下发到端侧，并将数据和画面进行绑定，同时在用户旋转手机时，通过对手势的监测选择对应的画面中心点，并进行画面的放大、缩小或平移。

2. 基于原始的大量数据样本，做算法模型训练，得到对于视频画面区域热度算法的模型。由于视频是一个连续过程，我们需要对镜头的切换做平滑处理，结合算法生成一个原始的算法数据。



3. 将算法数据和视频内容做关系绑定，并下发到端侧。这样在端侧就同时具备算法的数据和实际播放的视频数据。在播放进程中，我们需要获取旋转 – 陀螺仪传感器的输入，也会利用降噪算法过滤躁点，根据用户的旋转角度，结合当前视频画面，将算法数据和画面本身绑定，找到画面中心点，做相应处理，最终渲染到屏幕上。

以上是大致实现思路，在落地过程中，我们也面临不少挑战，最突出的是算法与传统图像处理算法不同。普通的图像处理多是基于单张图片，而视频本身是多帧的，而且每个视频帧间是连续性的。同时在识别过程中，尤其对于运动场景、切换镜头的场景，普通算法的识别焦点是存在偏差的，甚至识别不到，所以我们需要新的处理。



在算法设计上，采用镜头分割方式，区分不同的场景镜头，然后对于每个镜头，我们认为是画面是连续的。这部分，我们结合现有成熟算法，融入自己的技术探索。

首先，在看画面时，人眼睛会聚焦在人脸、人体，这些点的区域热度是比较高的，将些场景样本作为模型训练数据，同时视频本身还有部分字幕，也需要去除、识别和检测的处理。综合这一系列的检测内容，最终把一帧帧画面看成一个连续的轨迹，做聚类，形成一个角色或者是一个热度点的轨迹；集合多个镜头，形成一个视频区域热度算法的数据，然后下发到端侧。

其次，有了算法数据，在端侧更多是如何处理端上传感器，处理算法数据和视频之间的同步问题。

以上是我们现阶段的尝试，同步也在做其他尝试，在不远的未来也会逐步上线，大家很快就体验到。

三、基于 5G 的云渲染

在现有场景上，算法数据是基于原始视频进行识别，由于中间需要预生产过程，这就局限了它更多是在点播场景中。

如果不做预生产，而在端侧进行，则会产生识别的速度不够，效率底的问题，以及在不一样的交互时，处于实时性的诉求，本身对端侧算力是非常大挑战。结合 5G 的发展，我们设计出云渲染方案。首先看两个视频：

云渲染

6Dof 视频

（点击链接 httpsdeveloper.aliyun.comarticle765074，观看视频效果）

视频 3 在电脑上，可以认为它是一个云端主机，在云端是一个高清画质。而在手机端，用户真正看到画面，只是云端画面的一部分。为什么这样设计?

视频 4 是 6DoF 视频，用户可以通过手势旋转，从各角度看到不一样的视角。

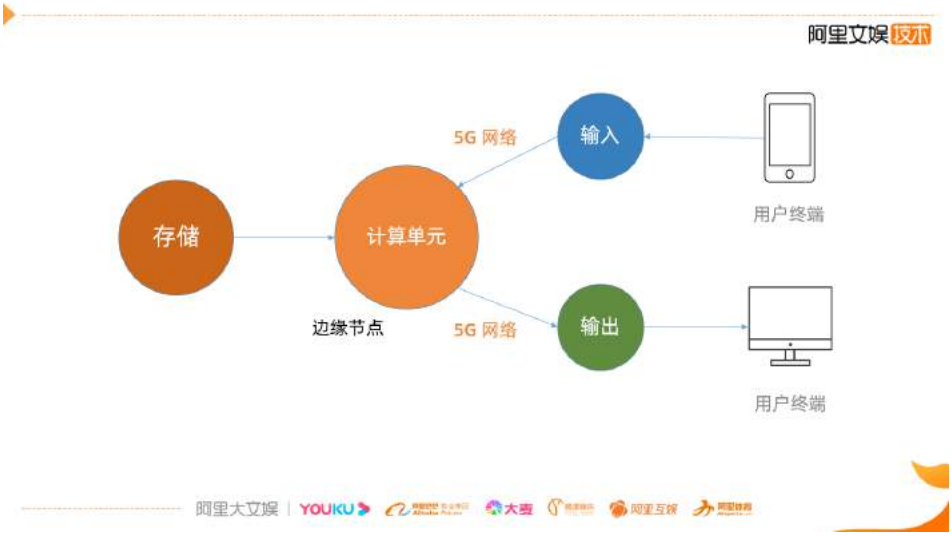
6DoF 视频的本质是，用户看到的某些角度的视频，其实是很多角度拼合的画面，用户在选择某一角度时，我们经过截取，提取其中两个画面，通过算法虚拟生成，一个用户观看角度的这么一个画面，然后下发到端侧。

6DoF 视频的某一帧，真实画面本身是非常大的画面，8k 甚至 11k。用户端看到是其中一部分，720p 或 1080p，其对应的 VR 场景也类似。

挑战是什么? 用户观看 VR 全景视频时，本质是 4k 甚至 8k 视频，但用户在每一个视角上看到的点，可能只有 720p 甚至更低。想看更高清的画质，就必须提升画面的大小。如果我们希望要看到 4k 画面，原始画面要达到 8k，甚至更高。

8K 画面下发到端侧是无法解决问题的。一是芯片的限制，其次还有电量、能耗

等。所以我们将终端计算能力放到一个强算力中心上，将用户终端设备变成三部分：手势输入、屏幕输出，计算单元放到远端计算服务器上，它的算力要数倍甚至是几十倍于端上。



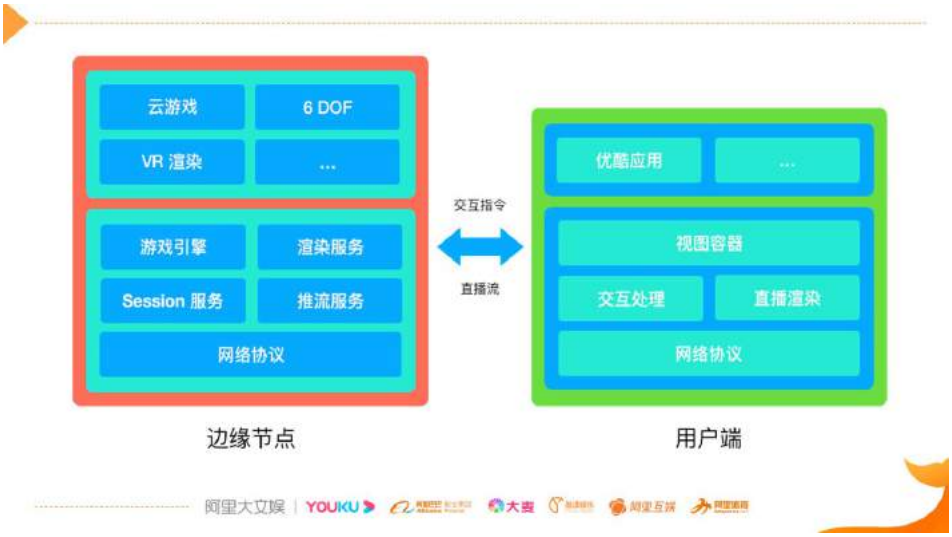
基于分布式的前提，输入、计算和输出的传输过程的耗时变短。考虑到未来 5G 网络、边缘计算的发展，在边缘节点和终端之间的传输速度，加上边缘计算节点的计算耗时，可能要比你在本机输入到本机芯片计算的耗时还要短。

所以，我们设计了一套分布式的云端渲染和实时计算方案。一方面解决交互的方式，大计算量的实时的数据场景。另一方面，借鉴在游戏领域（如云游戏）的思路，以下是设计模型：

1. 对于用户的手机终端、VR 设备或眼镜类各种设备，因为硬件在不断发展，它的算力会越来越强。但是个别设备算力还比较弱，所以我们希望有实时调度能力。算力强的设备，在端上做；算力弱的设备，在云上做。同时基于用户的手机电量等各方面场景，在边端体系上有一个调度能力。用户端的一个播放行为，其实是从媒资的存储到转码、CDN 分发，CDN 节点，通过分发服务到手机终端，当用户点击视频，通过对应的时间节点拉取对应的云端视频数据。
2. 在云渲染链路上，我们希望用户是通过调度的操作，决定计算逻辑是在端上还是边缘节点上。如果在边缘节点，通过边缘节点去访问中心节点，拉取到数据。当用户再次操作时，通过边缘节点进行相应的交互处理，再下发到端侧。这样从边缘节点到播放终端，是点对点的实时传输的操作。

细化云渲染的整体设计，我将它分为五个部分：边缘服务框架、网络协议、端侧交互引擎、边端调度系统、应用开发工具链。其中边缘服务框架、网络协议、端侧交互引擎如下图所示，分别承担着边缘节点的框架服务能力、网络通信的协议处理、以及终端的交互、渲染引擎。而边端调度系统如上所说，主要是根据用户终端、边缘节点算力等情况合理调度用户的渲染服务是应该在终端处理还是到边缘节点处理。而基

于此，我们可以看到，很大程度上服务程序是需要在多平台基础上运行的，所以相应的开发工具链（开发调试 IDE、服务部署发布系统等）也是很重要的部分。



在边缘服务上，我们希望搭建一套基础框架，不仅承载现有的渲染服务，未来也可以部署游戏引擎来实现云游戏的服务。由于单个边缘服务节点需要服务多个终端设备，推拉流服务的用户 session 管理很重要，并且低延时的推流处理、高性能的渲染服务等都是我们需要突破的重点。同时，由于我们定义的很多场景是基于实时计算和强交互的模式，更像是游戏，上行的数据以操作指令、文本等为主，下行则主要是流媒体数据、算法数据等，而且考虑到时延等问题，优选基于 UDP 构建的传输协议，同时考虑到网络穿透率的问题，基于 TCP 的方案会作为基础的兜底策略。而在端侧，重点是低延时的直播播放器，网络协议的客户端实现以及用户上行的指令处理等。

四、新交互的未来是什么样子?

始于播放新交互，结合 5G 和边缘计算，面向云渲染。基于这个链路，未来我们希望的播放新交互是什么样子?

首先，在交互能力上，我们已经将算法和内容做结合，视频内容本质上是导演、演员基于剧情，向用户传递信息。用户观影过程中，是不是可以跟导演、演员或内容之间有联动交互。

其次，如何将这两者之间的信息拉通？就是通过算法结合内容做识别，算法本身去识别内容，再将识别出的内容跟用户看到的内容，在信息上更贴合用户偏好，将更多主动权交给用户，给用户更沉浸式的观影体验。同时因为这种交互的模式，对算法对算力有更高要求，借由 5G 和边缘计算的发展，打造一个环形体系，实现播放新交互的体系化形态。

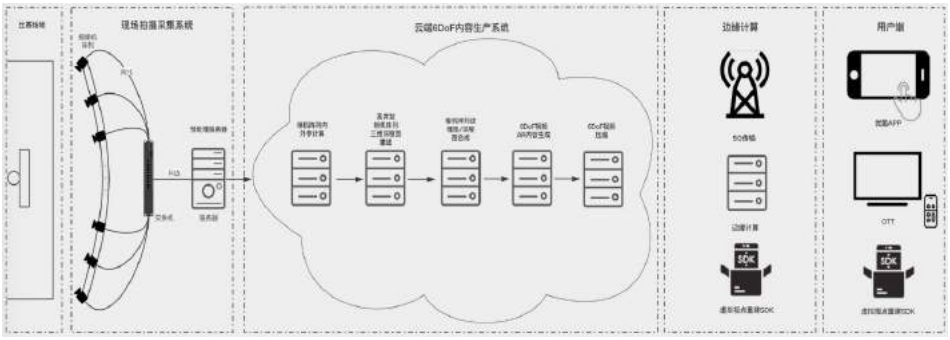
以上就是我们对于未来在播放和交互领域的思考。

AI 篇

概况　阿里文娱智能算法的新应用

作者 | 阿里文娱资深算法专家　胡尧

一、文娱消费新体验 – Free Viewpoint Video



面向文娱消费新体验，文娱算法团队基于整体的视频生产、播放、交互式体验等环节做了非常多的技术探索，在视频子弹时间的基础上进一步延展，延伸成更加经济通用的 Free Viewpoint Video 技术，构建完善的现场 – 云 – 边 – 端的技术链路。

今年优酷与 CBA 达成全方位的合作，在新赛季首次落地互动 FVV 体验，变革传统体育赛事的观看体验。我们还主导建设 FVV 视频技术国家标准，同时承担了国家“科技冬奥”“冰雪项目交互式多维度观赛体验技术与系统”项目，让更多普通用户享受到新一代观看体验。

二、视频消费商业化新模式 – 创意广告

在视频消费商业化新模式上，重点在创意广告领域，打造动效酷炫、更原生、可互动的广告形态，通过规模化实现中长尾内容的变现。

具体是构建了综合的三维环境感知和理解算法，在多场景多模态的视频内容理解、云端结合的 CG 渲染引擎等领域都有深度探索，并且构建云端结合的 c c 级渲染引擎，实现 AI 两大平台（视频打点平台、视频特效生产平台）、一个中心的战略（创意广告制作中心）。

三、基础素材自动化生产与检索

在视频理解领域的另一个重要应用，就是搭建了视频创作新平台。

首先，我们实现了基础素材的自动化生产与搜索，将传统的静态人物动作和场景实现升级到 shot 级别，支持动态素材的检测与识别。

同时，实现精细化抠图，包括对动态人物和物体场景的提取，使这些素材可以被创作者更快速的应用到新的创作环境中。通过交互式的精细化分割算法，动态地将经

典视频中的素材提取出来，实现简易化绿幕的效果。

当大量的素材被生产出来，我们同时提供基于准素材级别的智能化检索系统，用户只需要通过语义的文本或语音输入，就能实现对整个素材库的检索。例如用户搜索“吴倩拥抱”，系统就会呈现出整个《冰糖炖雪梨》中有吴倩拥抱的场景。

四、封面图自动化生产

另外，我们实现了封面图自动化生产。基于主要人物、场景、美学评级、元素多样性等方面生成不同维度的封面图，并提供智能裁剪服务，满足 16:9、4:3 或者 3:4 等各种场景需求。同时在某些场景中实现动图的自动化生产，即实现千人千面的内容 + 素材的统一个性化推荐，助力运营分发的提效。

五、模板式视频半自动化生产

鉴于优酷有海量的 IP 版权内容，我们研发了一系列的剪辑合成技术，自动对视频的故事线、内容模板进行提取，并在此基础上在海量视频中进行智能化的二次创作，实现如节目卡点剪辑、Video Highlight & Summary 技术生成的前情提要等产

品。同时具备视频的形态转换技术，将横版的视频通过 AI 算法，识别显著性主体区域并进行美学评判，实现竖版视频的自动化生产。

这些技术能够有效的为商业化提供更多素材，同时为 B 端提供更多能力。

在这个基础上，我们才能实现基于元素级的视频深度理解技术，我们将传统的基于用户行为的内容分发体系和基于视频内容理解的视频内容分发体系进行了有效结合，实现了群体智慧和计算机视觉在美学和 AI 上的融合，实现了从整个封面图内容的原数据分析，到整个用户行为偏好的判断，实现千人千面的内容加素材的个性化推荐，有效提升整个业务场的分发效率。

第一章　XR 技术在优酷的应用

作者 | 阿里文娱高级算法专家　方如

大家都看过科幻电影吧，像《头号玩家》、《美国队长》、《银河护卫队》，这些科幻电影中都出现过 AR/VR 的镜头。以《头号玩家》为例，主角来到博物馆，能够实时的、多角度地去浏览资料，这里就用到了 volumetric video 技术，它是一种 VR 技术，就是在被摄物周围放一圈摄像头，采集的视频经过合成加工就可以无缝地切换观看了。与电影中的拍摄特技不同，随着 5G 和 AI 的加速落地，在视频生产和播放环节，越来越多的融入 AR、VR 的相关技术。可以说今天的科幻电影是明天的科学事实。

那么优酷在这一领域是如何实践的呢？且看阿里文娱高级算法专家方如在 GMIC Live 2020 智慧文娱技术专场中的分享，主要从四方面展开：

- 一是 XR 与视频的结合策略。
- 二是 XR-Video 技术特点。
- 三是 XR-Video 智能创意平台及其应用。
- 四是 XR-Video 未来展望。

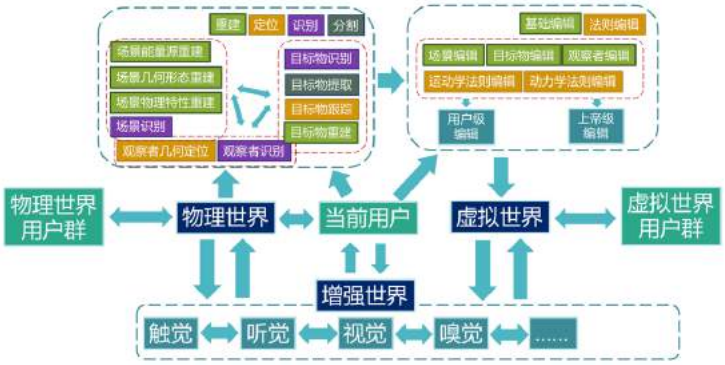
一、XR 与视频的结合策略

什么是 XR？ XR 包括 VR、AR 和 MR。

什么是 XR?

VR 是 Virtual Reality 虚拟现实，是计算机模拟出的世界，给人一种沉浸感。AR 是 Augmented Reality，将虚拟物体放在真实世界中，但与真实环境不能交互。MR 是 Mixed Reality 混合现实，将虚拟现实和增强现实进行融合。在 MR 世界中，真实实体和数据实体是同时存在，可实时交互。

XR 系统组成



核心要素：世界感知、世界编辑和交互 来自源方

XR 系统的核心要素是世界感知、世界编辑和交互。1) 世界感知是对物理世界的一个逆向工程，将物理世界数字化；2) 世界编辑是一个正向工程，打造虚拟世界。将物理世界和虚拟世界进行融合，就打造出了一个增强世界 ;3) 交互包括用户和世界之间，用户和用户之间的交互。可以看出，XR 是一个极其庞大的系统。

XR 与 视频结合策略

优酷在不断探索，将 XR 理念和手段与视频结合形成 XR-Video 技术。它具备以下特点：

1）虚拟信息与视频富有创意的植入。

2）时空多纬度的合成。

3）实时的交互性。

二、XR-Video 技术特点

一是虚拟内容与视频富有创意的植入。

1. 植入形式

虚拟内容的植入形式是非常丰富多彩的，我们创造了高光时刻、悦享时刻、移花接木、无中生有、动态混合现实等十几种的酷炫特效。比较典型的移花接木，就是找到物体的平面后替换原平面中的内容；动态混合现实，是在视频中植入运动的虚拟内容。

移花接木视频 （点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

动态混合现实视频 （点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

2. 植入内容

植入的内容从哪里来呢？有两个来源，一是从素材和特效库里提取；二是在原视频上利用 AI 算法智能的生成特效。

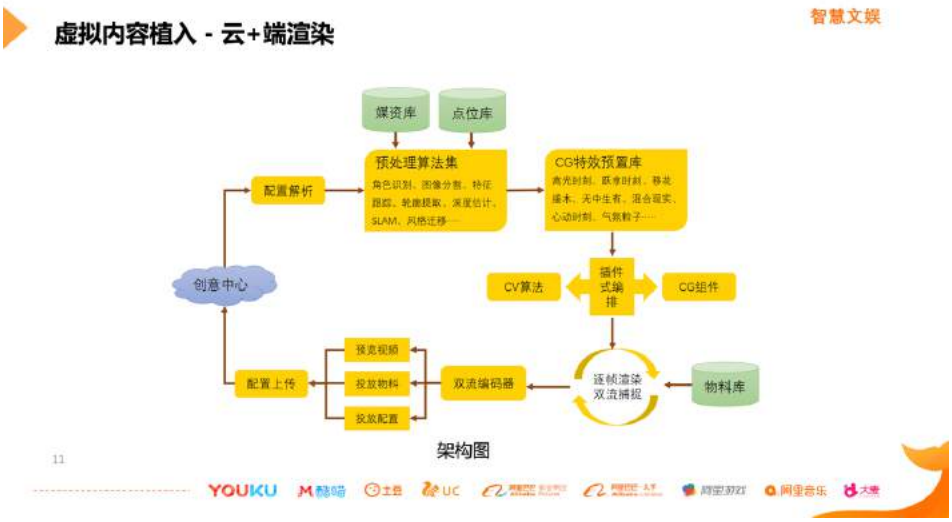
人物复刻视频

高光时刻视频

子弹时间视频

（点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

将视频中的人物图像分割出来，进行复刻，生成人物复刻特效，如左侧视频所示。通过人物的检测分割形成轮廓，粒子绕着轮廓进行环绕，形成了高光时刻的特效，如右上视频所示。我们与阿里体育合作的子弹时间，通过 CV 算法智能识别出球员、弹跳高度等等，这些数据生动形象地在 6DoF 视频中展现出来，右下视频所示。

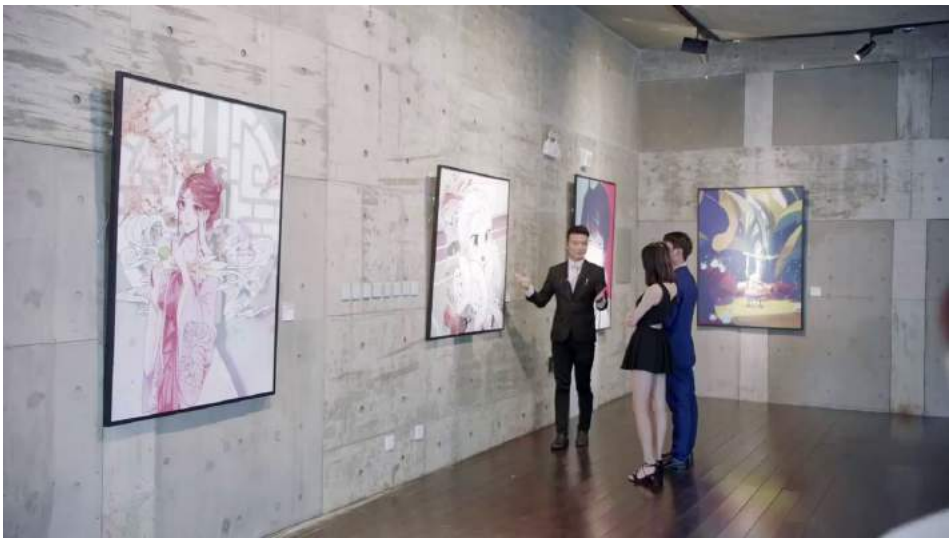


3. 内容呈现方式：云端渲染

支撑特效制作的是大千云端渲染引擎，它支持自动化和规模化。以植入广告为例，从创意中心下单，经过特效引擎制作和投放，实现了近自动化，保证了特效广告上线的及时性。与传统特效制作方法相比，我们有两大优势，一是传统的影视包装技术比较匮乏，难以与 CV 算法结合进行创新，而我们采用开放式 CG 方案，将物理计算、粒子系统、光影渲染等 CG 技术进行插件式配置，灵活地与 CV 算法结合创新；二是传统方法特效制作周期长，人工介入流程多，动态性差，我们采用了实时渲染和双流捕捉技术，大大提升制作效率。

除了自动化和规模化，植入渲染的品质和细节是我们的关注点。以移花接木植入渲染效果优化演进为例，美学自然的融合，实现特效与原场景的自然的 XR 植入，是

我们追求的目标。视频中的待植入区域通常是存在运动、形变的。如下视频所示相框区域跟随镜头移动，且因透视原因存在形变。



移动和透视形变视频 （点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

若简单地采用 Mesh 来复现点位结构信息，在植入时会出现纹理的闪动和边缘的锯齿，因此我们优化了纹理平滑和边缘的抗锯齿工作，拉通抗锯齿和浮点插值渲染，使植入初步达标。在这基础上，下一步工作是把植入位的图像风格迁移到待植入的素材图像，使得植入后的素材区域和原始视频的整体图像风格一致。为此我们引入深度学习方法结合 Wavelet Transforms，实现了植入后的素材自然，无违和感。

USDF 处理前后对比

上图的差异可能较小，视频放大以后，尤其到大屏播放能看到清晰的毛刺。通过距离、UV 梯度对边缘进行柔化、对纹理信息进行微调，解决毛刺等问题，让整体植

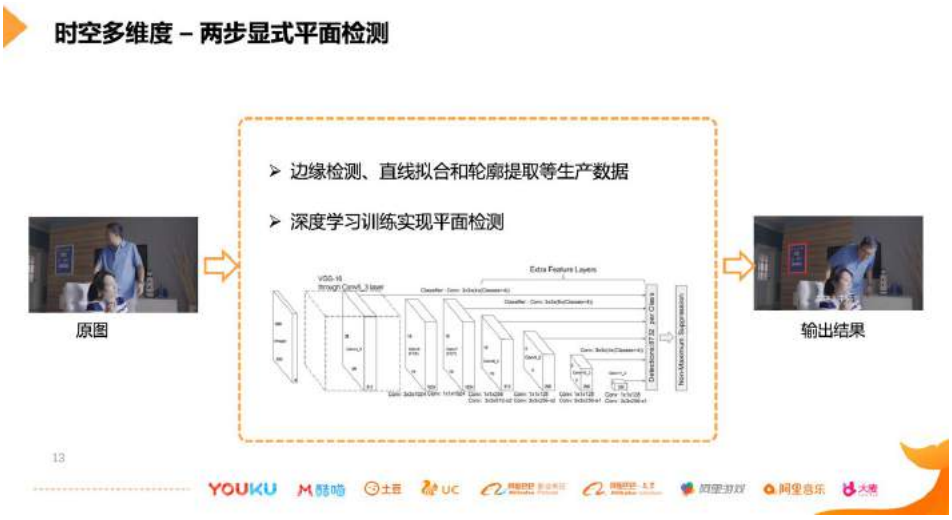
入区域更柔和自然。在采用了 USDF 着色算法，经过风格迁移处理后，《长安十二时辰》海报非常自然地融入到视频中，后续进行了多虚拟相机分层，Blend 二次处理，提升叠加的易用性，得到更佳效果。

移花接木结合风格迁移与 USDF 处理的植入视频

（点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

二是时空多维度。

“时”是对视频进行打点，具体是通过对物体和场景等的识别，理解视频内容，确定植入时间，目前已实现帧级别。在“时”的基础上，我们加入了“空”的感知和理解，确定植入的空间位置。以典型的移花接木为例，需要平面检测和平面追踪能力。



显式平面检测包括对海报和平面等检测。采用了两步显式平面检测方式。通过对边缘检测、直线拟合和轮廓提取等手段生产出初步的平面数据，辅助人工标注微调。这些数据再通过深度学习进行训练，进一步提升准召率。

隐式平面检测包括墙面、桌面、楼面等。通过隐式平面检测，进一步扩大植入场景。采用传统的方法 SFM 三维重建，从视频序列中计算相机的 pose 恢复稀疏点云，再通过 CMVS/PMVS 重建稠密三维点云拟合平面。

但是传统方法在视频中有运动物体，当特征点较少时，效果非常不好。在这种情况下，推出了深度学习方案。使用 CNN 估计图像深度信息，重建 3D 坐标。通过图像超像素分割获取 cluster 处理，判断共面进行隐式平面（墙面等）检测。

平面追踪算法是移花接木的核心技术之一，分为三大类：基于区域、KCF、基于特征点。我们采用的是基于特征点的方法，利用深度学习进行特征点和描述子的自适应联合学习。传统特征点有 SIFT、SURF、KAZE、AKAZE、BRISK 和 ORB 等，Learning-based 方法，例如：D2-Net、R2D2、LF-Net、SuperPoint 和 UnSuperPoint 发展迅速。基于深度学习的特征点提取是今后大趋势，它也是 SLAM、image-based localization 等应用的基础能力。下图是 DOG 和 UnSuperPoint 特征点提取效果对比。经过对比，深度学习方案从 reliability 和 repeatability 方面优于传统方案。我们采用 UnSuperPoint 方案进行特征点的提取和描述子的计算。

平面追踪的四个改进方向有：（1）利用深度学习进行特征点和描述子的自适应联合学习（2）可靠的特征点提取后，引入图模型和图匹配，（3）结合 H 矩阵平滑提升单应性矩阵的准确性。（4）融合多种网络。采用的二阶段高精度平面追踪，结合 attention 机制，对人和物体遮挡引入的噪声像素进行屏蔽，实现了在运动且遮挡情况下的稳定追踪，且优于 AE 追踪的结果，参看如下对比视频。

大千平面追踪与 AE 的效果对比视频

（点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

曲面追踪可以进一步扩大植入的应用场景。从特征点计算、特征点匹配和筛选，实现三角面片网格化。在此基础上添加植入元素转换成 UV 贴图，然后进行特效渲染。扭曲运动物体表面的追踪后进行文字、Logo 或动画等植入。下面段视频显示了跟踪和植入的效果，植入生动自然。

《大千植入行云流水》演示视频

（点击链接 https://developer.aliyun.com/article/765084，观看视频效果）

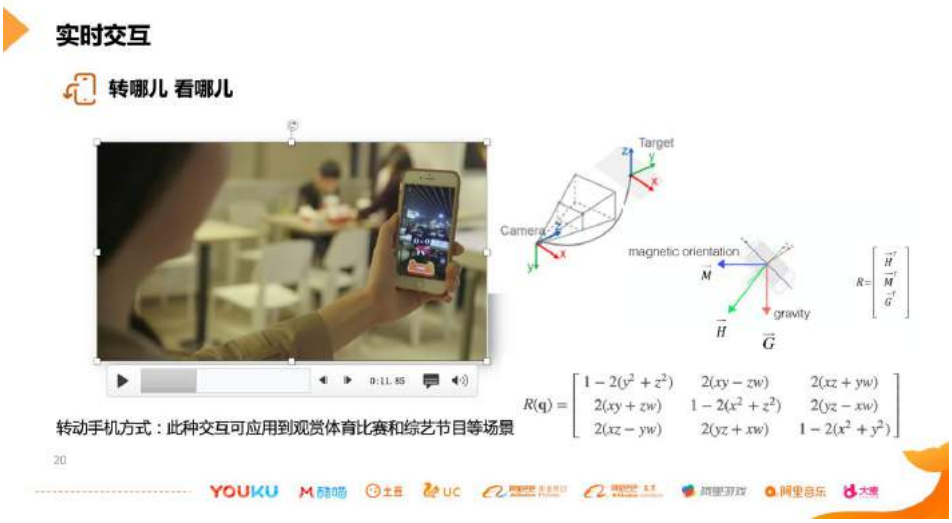
三是实时交互。

交互从简单的人面对屏幕观看视频发展到将 2D/3D 信息融合于周围的空间与对象中，不再与视频内容脱离，而是和人们的当前视频自然而然地成为一体。交互的动作除了以往的按键或者触屏，可以扩展到头部、眼部、表情、手势和语音等，从位置扩展到原有视频某个空间。分享一下实践的三种交互方式。

“点哪儿活哪儿”

实际上就是一个 3D 模型交互。例如在视频广告中，我们可以在出现保时捷品牌汽车的点位进行预埋点，通过特效触发召唤出汽车模型，用户可以与汽车模型进行三维触控互动，模型可动态展现品牌汽车的各个角度以及开关门、开关灯等各种行车效果，这种 3D 互动式广告可以大幅增强广告的品牌感知度和认可度。

“转哪儿看哪儿”



转动手机进行交互。例如 AR 捉猫猫是类似 Pokemon Go 的游戏，是 LBS+AR 技术的一种成功运用。游戏活动期间总 PV 十几亿，日均 UV 三千多万，支持星巴克、KFC、苏宁易购等 60 多款品牌猫，是那年最火的双 11 预热互动活动。在这个

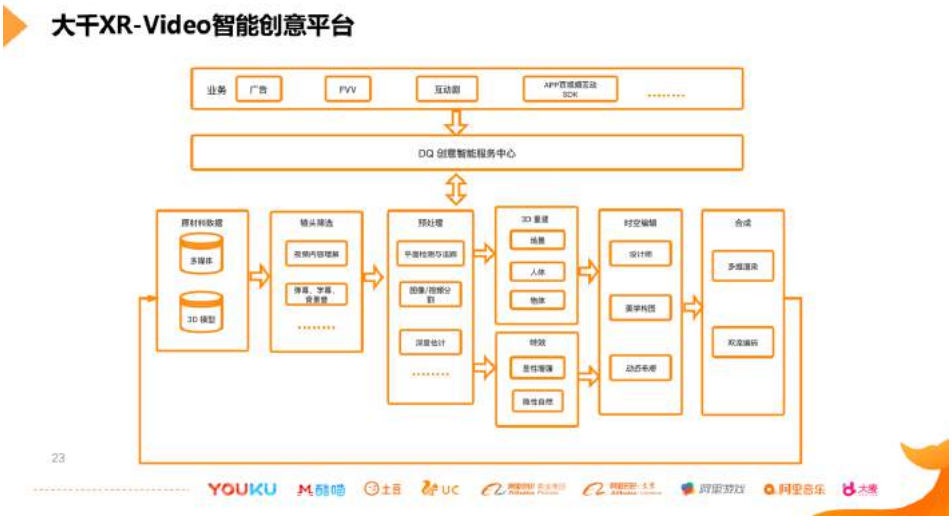
游戏中，主要解决的一个问题是，通过手机的加速度计、陀螺仪，磁力计和 GPS 信息，实时计算出 3D 模型在屏幕上的显示位置，给用户一种该 3D 模型（例如星巴克猫）就在其真实世界周围的某个方位上的“错觉”。这个“错觉”的视线方向通常表示为一个旋转矩阵。这种转动手机的玩法，还可运用到观赏体育比赛和综艺节目等场景中。

“看哪儿买哪儿”

在 VR/AR 中通过空间定位，人置身其中，参与其中的互动，犹如身临其境一般。“看哪儿买哪儿”实际上就是基于视线的交互，它是一种更自然的人机交互方式。视频展示了沉浸式购物全景视频，用 VR 手机盒子体验的购物应用，带你穿越到世界各地的商场购物，遇到喜欢的商品，用户盯住触发按钮就可下单购买。

三、智能创意平台及其应用

XR-Video 创意平台意在打造生产和消费的生态系统。从获取原材料开始，然后进行各种筛选。原材料包括有多媒体和 3D 模型。筛选方式有视频内容、弹幕、字幕和背景音等的理解。通过深度估计、平面检测和追踪、图像和视频分割等方法进行预处理，然后进行人体、物体和场景的 3D 重建，以显式和隐式的方式叠加特效，进行时空编辑，最后通过多维渲染、双流编码的方法进行合成并生成一个特效视频。生成的特效视频一条路是返给创意智能服务中心。创意智能服务中心担任与外部应用对接的角色，通过它服务于广告、自由视点视频、互动剧、APP 页互动等应用。另一条路是返回给原材料库，形成闭环实现良性的循环。



大千 XR-Video 智能创意平台框图

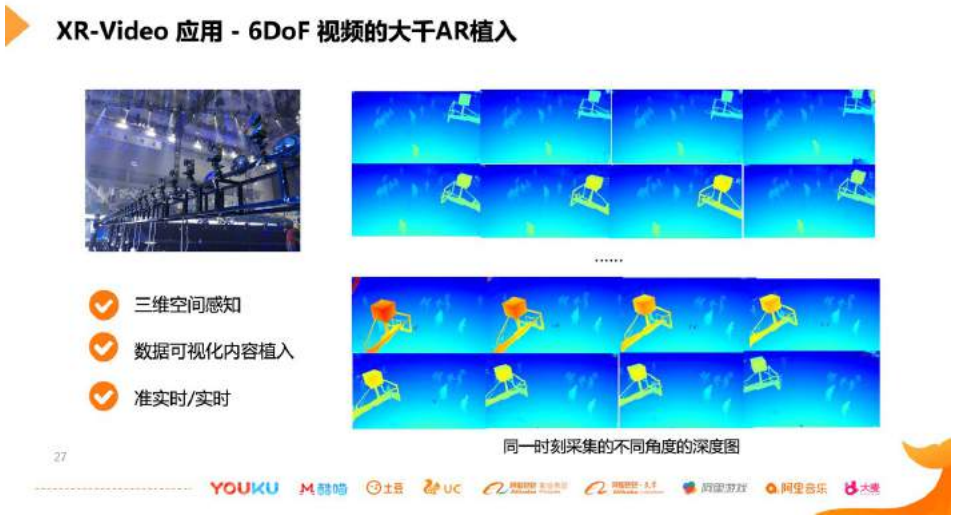
应用之一：大千植入特效广告

传统内生广告有压屏条和创意中插。压屏条样式呆板，俗称“牛皮癣”；创意中插要前期制作，成本高；所以我们创造了特效广告形式。曼秀雷敦、良品铺子、OPPO 和哈弗等多家广告主上线尝试了这种新型的广告形式。

大千创意广告，替代了枯燥的硬浮层广告，创造出全新的广告观感；是一种不打扰观影的软植入，解决了会员用户和广告客户间的利益矛盾。其涉及的技术点有三维环境感知、HDR 光照估计、特效 3G 渲染引擎，多维度多模态视频解构打点以及严格的帧同步。

严格的帧同步是在保证播放原视频的同时，还要保证动态渲染广告的视觉效果，支持个性化更新。在千元机上实现严格帧同步挑战挺大。 帧同步双流渲染技术经过了多次迭代，直接叠加带透明通道的视频会存在兼容性问题，因此我们提出了滤色 Key 方案（性能消耗较大）、WebP 渲染方案（内存占用较大），逐步演变到了双流掩码方案（性能、资源占用情况均较佳），最后通过 pts 基准合流渲染的方法达到了严格的帧同步，至此特效广告与视频资源达到了解耦 + 同步的两全其美效果。

应用之二：6DoF 视频的大千 AR 植入



6DoF 视频的大千 AR 植入的实现方式是在体育场馆里布置一圈摄像头，摄像头采集 RGB 和深度图。图中显示了同一时刻在不同角度拍摄的深度图。通过三维感知计算，准实时和实时生成更多虚拟视点的图像及其相机位姿，叠加 3D 动效，实现数据内容可视化的植入。

在 6DoF 视频中植入，一个重要功能是热区图。通过三维重建方法进行三维场地标定，利用深度图去篮筐遮挡，结合人体姿态估计与分割去人体遮挡，最后实现了根据虚拟试点相机姿态进行热力图的渲染生成。

铭牌组件植入演示视频

6DoF 视频 AR 植入的另一个功能是铭牌组件。需要解决的问题是有球员识别和跟踪、篮球识别、手和脚识别和定位以及人体的三维建模。通过 3D 建模去计算高度等等。越来越多智能生产的数据通过这种方法可视化，而且还可以动态地植入广告。

四、XR 研发方向

XR 的研发方向，就是更沉浸、更准确、更有趣的互动式植入。

第一个方向，基于 2D 与 3D 结合的三维感知技术。具体来讲，利用三维点云语义分割建立人与物、物与物的相对关系；利用人物遮罩与 3D 景深结合，处理遮挡问题和场景切换问题；还有利用空间位置结合手势等识别的进行 3D 交互。

第二个方向是光照估计。左图为原始图，右图为光照估计后的植入效果图。可以看出，光照估计后植入物体的阴影是非常自然的。我们采用基于深度学习“场景光照特征”识别算法，智能感知视频画面中的光源方向及照度分布，保证虚实场景视效的一致性。

一直坚信 XR 是改变人机交互的终极形态。但由于硬件发展还存在一定的问题，XR 眼镜的续航、重量和效果等需进一步提升。这段硬件改进的时间，正是积累 AI 算法和 XR 应用软件技术的时机。AI 是 XR 应用的基石，通过 AI、XR 与教育、培训和游戏等场景结合在实践中摸索，等硬件成熟后定能实现厚积薄发。

三维空间感知和理解是 XR 技术的核心之一。传统的 SLAM 技术关于测量、几何的方法虽然已经比较成熟，但面临着发展的瓶颈。如果要大发展的话，需要结合多传感器和深度学习的方法。深度学习是工具，SLAM 是应用的关键技术。除了深度学习外，在 SLAM 技术中加入仿照人类对环境的感知能力和特殊的先验约束等手段提升定位的速度和精度。有了准确的空间位置感知，与动作捕捉和语音交互等结合，实现自然的人机交互。

AI 是人机交互和人物景理解的基础，而 5G、边缘计算强有力支撑了 XR 所需要

的大数据传输，他们结合起来催生了 XR 的发展。了解视频编解码和边缘计算等技术原理，关注这些技术的发展趋势，对设计和实施 XR 应用系统会有很大帮助。

XR-Video 还在探索中，它会继续在创意广告、互动视频和视频制作等领域上施展拳脚。

第二章　竖屏看热剧！阿里文娱横转竖技术实践

作者 | 阿里文娱算法专家　闵公

常见的机器视觉问题，诸如目标检测、主体标定、目标追踪、视频增强等作为独立技术问题来求解，是不是有些枯燥？在文娱产业中，如何将这些视觉技术进行创新和组合形成完整技术栈，对海量横屏播放的影视剧和短视频自动转换成竖版播放的视频？

且看阿里文娱摩酷实验室的算法专家闵公在 GMIC Live 2020 智慧文娱技术专场中的分享，主要介绍如何“基于机器视觉算法自动化”将海量横版长剧集转换竖版视频，包括横版视频的主体自动选择算法，镜头平滑能力等，希望对大家在视觉算法如何运用在文娱行业中有所启发。

核心技术内容包括：

1）视频横转竖技术链路搭建

2）复杂环境下主体自动标定模型

3）shot 镜头平滑和标定追踪交互机制

4）视频裁剪导致降质条件下的画面恢复

一、横屏转竖屏的视频裁剪的行业需求

首先，站在海量内容消费者的角度来看，90% 以上的视频内容消费者会选择单手竖持手机，同时也有 50% 以上的用户会选择将屏幕进行竖向的锁定浏览。同时视频内容消费者倾向于将视觉聚焦在焦点主体内容，而不是背景上。

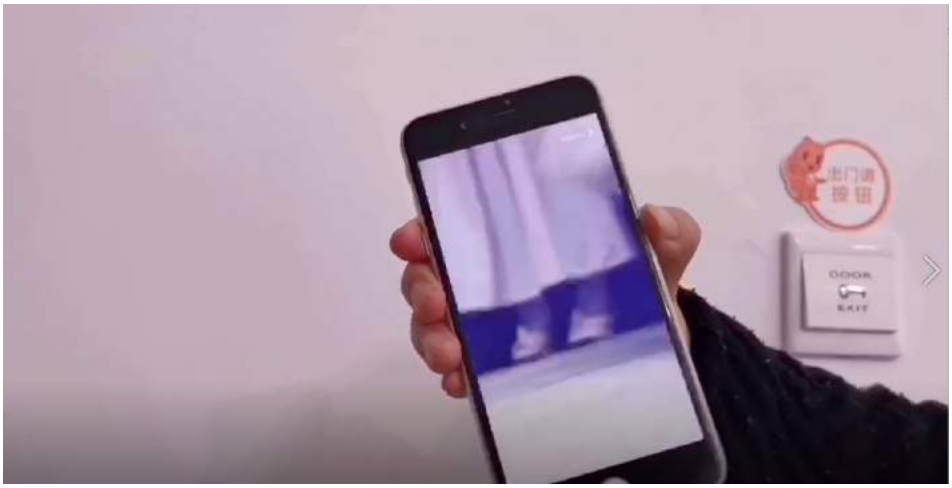
其次，站在内容生产侧，大剧制作有 95% 以上是横屏拍摄，站在 PGC 短内容的制作上，像优酷全娱乐、体育等，多是以横屏内容制作为主。如果通过人工将横版拍摄的视频进行剪辑转竖，效率低，效果不可控。

二、横屏转竖屏的产品落地化进程

优酷人工智能平台推出自研的视频自动化的横屏转竖屏技术，应用于视频二次生产和智能封面图生成业务中，目前该技术已经覆盖优酷的 OGC 剪辑，海量 UPGC 竖版短小视频生产，智能封面图生产，同时输出给阿里云，服务于文娱企业客户。

三、横屏转竖屏的视频技术链路

智能裁剪技术主要应用于以多人或者单人为主体的影视剧场景，我们将目标检测，跟踪，识别等技术进行创新和结合，开发了完整的视频智能裁剪技术链路，面对实际业务中的主体标定，视频帧间抖动，视频黑边填充等问题针对性的研发了算法解决方案，可以根据不同的业务场景将各算法可插拔的配置进主裁剪 pipeline 中，阿里文娱视频智能裁剪技术的研发给内容行业的素材自动化制作，剪辑作品的视觉效果和制作成本降低等方面都带来了大幅度的提升。



（点击链接 https://developer.aliyun.com/article/765081，观看视频效果）

在视频智能裁剪技术链路中，我们研发了前处理模块（包含镜头切分，画面尺寸判定，黑边检测裁剪等），主体选择模块，主体追踪模块和后处理模块（包含画质增强，字幕 /logo 检测，画面内容修补等），下面分别介绍四个模块。

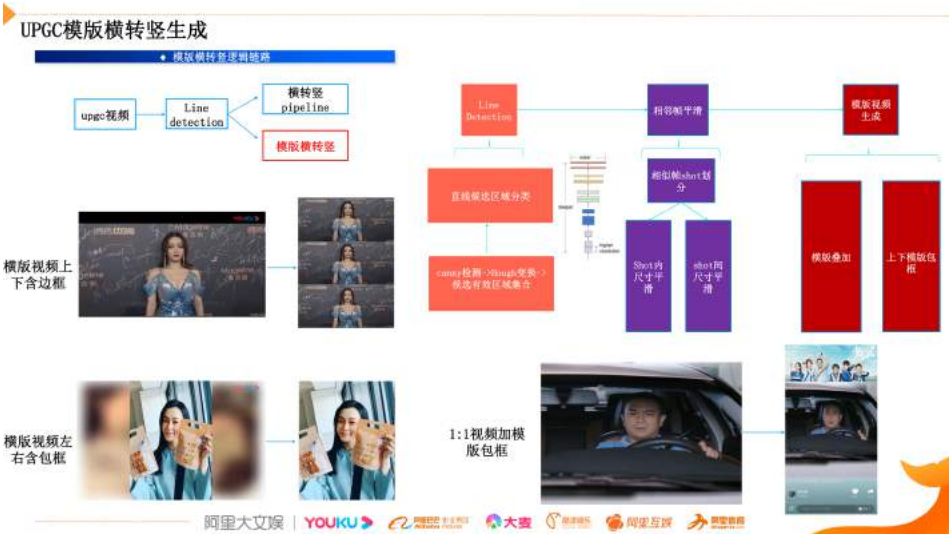
四、视频裁剪核心研发模块

（一）前处理模块

前处理模块包括分镜边界检测模型，画面尺寸判定算法，黑边检测与剪裁算法等三个模块，其中分镜边界检测模型根据视频画面将视频分成多个镜头片段，针对渐变过渡的镜头，采用视频帧的表征向量逐帧计算相似度进行精细切分。

画面尺寸判定算法使得裁剪可以在不同的画面尺寸中进行自动选择，包括（宽：高）16:9，4:3，1:1，3:4，9:16 等任意尺寸，通过对视频帧进行抽样后根据目标的显著性和运动特性计算得出显著区域的大小进行剪裁尺寸适配。

由于大量 upgc 横版和竖版视频存在上下黑边填充现象，但上下黑边在自动裁剪后会严重影响用户体验。因此我们使用霍夫变换和直线分类来解决黑边检测与剪裁的问题，并且根据剪裁后的不同尺寸自适应的选择三层重叠样式或者 1:1 加包框样式进行视频再生产。



（二）主体标定模块

我们根据人工标注的影剧综主体 GT 数据，设计了主体自动选择模型对视频帧中的主体进行自动标定。主体自动选择模型将视频中的人脸，人体 bbox，显著性区域，图像清晰度等候选区域进行 roi align 对齐后，通过深度卷积网络进行最佳主体选择模型的训练，通过和显著性模型以及注视预测模型进行 max-IOU 指标对比，我们提出的主体自动选择模型表现 SOTA。

同时我们将主体选择模型应用于复杂环境下的场景（如动物世界，大型晚会，新闻联播等）下进行效果测试，裁剪后的竖版视频效果符合预期，从而验证了我们提出的主体选择模型具备的泛化能力。

（点击链接 https://developer.aliyun.com/article/765081，观看视频效果）

在主体数据标注的过程中，我们制定了一套主体选择标注标准，包括主体中心化，主体 max 尺寸、主体尺寸比例，主体的姿态以及主体稳定性等。完成了主体图像数据集共 9.5k 的标注，视频数据集 125 个视频，共 13.2 万帧的标注。针对视频帧存在的多主体和人工标注的抖动问题，我们引入了 reid 和平滑滤波来为辅助解决上述两个问题。

（三）主体追踪模块

主体追踪模块包括目标追踪算法，镜头平滑算法，主体标定和主体追踪交互机制。通过对多个物体运行多次 SOT 追踪得到关键帧后续相邻帧中主体目标对应的位置，形成连续视频帧的镜头标定结果。我们在**追踪模块中引入 backward tracking 策略，将短时 track 能力扩展为长时跟踪，并进行了 local-to-global search based tracking，以此来降低追踪模块和主体标定模型的交互次数和计算时间。同时针对主体切分比例采取了黄金分割比例来提升美学观感。**

由于目标追踪算法得到的镜头剪裁位置并不是平滑渐变的，这导致画面抖动，引起用户观看眩晕等较差体验，因此通过时间序列离群点检测和 Kalman filter 等技术，将异常定位点 t 进行平滑，解决了裁剪后视频帧间抖动问题，抖动幅度 Jitter Degree 得到了显著性的降低，人工评估视频帧后观感流畅。同时通过主体标定和主体追踪交互机制，保证了主体目标在镜头切换情况下的镜头内容连续性。

（四）后处理模块

针对视频剪裁后的视频画质问题，我们开发了后处理模块（包含画质增强，字幕/logo 检测，画面内容修补等），主要解决剪裁边界可能的 logo/ 字幕截断问题和裁剪后主体相对放大和编码导致的分辨率降低问题。其中我们针对性的设计了去噪、超分辨率模型，对裁剪后的降质视频进行画质提升，在超分模型研发中，我们在训练数据增强上采用自适应采样算法（如下图所示，**红色 bbox 由随机采样得到，绿色 bbox 由自适应采样得到**）使得采样得到的图像 patch 集中在纹理细节丰富的区域，在模型设计上，采用了 multi-term loss 和 multi-branch module 的结构进行模型训练，最终超分模型在技术指标 psnr 和人工背对背打分上都得到了显著提升。

结束语

视频智能裁剪技术生产的视频和封面图广泛应用于优酷的各个场景，并得到了业务方和阿里云客户的一致认可，我们对视频智能裁剪算法栈进行了整体性能优化，达到处理时间仅 1:2 视频时长，目前该技术累计对优酷综艺：演技派，这就是街舞，这就是灌篮；优酷剧集：陆战之王，天雷一部之春花秋月，微微一笑很倾城等百部 OGC 进行裁剪服务，裁剪后的竖版视频用于抖音，微博等外渠宣发和站内投放，同

时主体标定算法服务于搜索双列封面图生产，镜头平滑算法服务于弹幕人脸项目，视频裁剪算法已经部署在阿里云上，由于目前行业内竞品尚无成熟技术方案，已经通过申报《基于主体目标标定与追踪的视频智能剪裁技术》,《基于智能画面分析和多层级主体目标标定的图像智能剪裁技术》专利的方式来保障该产品技术的竞争优势，期待阿里文娱视频裁剪技术为中国的视频娱乐行业创造更大价值。同时感谢 AZFT 计算机视觉与分析实验室的朱建科老师在项目过程中的技术指导和大力支持。

第三章　让用户快速找到内容的搜索算法实践

作者 | 阿里文娱高级算法专家　若仁

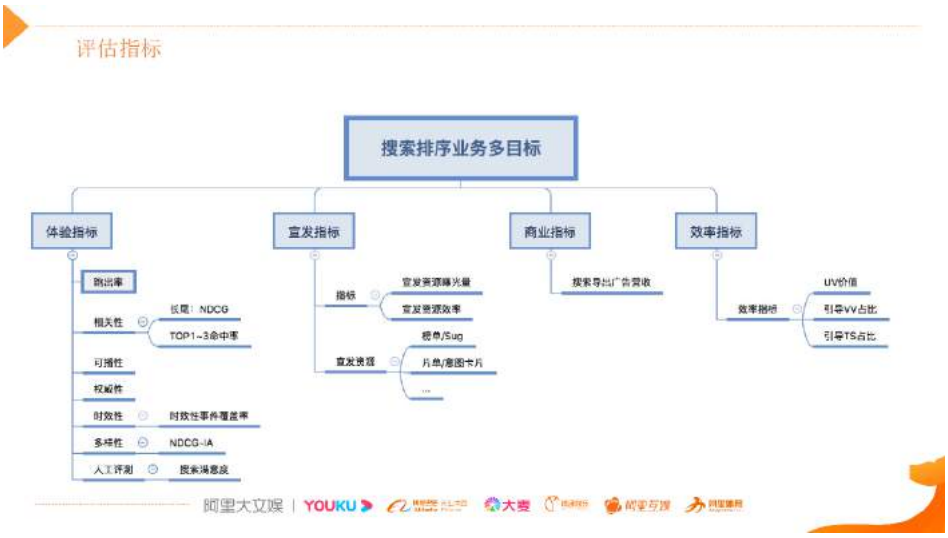
视频搜索是涉及信息检索，自然语言处理 (NLP)，机器学习、计算机视觉 (CV) 等多领域的综合应用场景，随着深度学习在这些领域的长足进展以及用户对视频生产和消费的广泛需求，视频搜索技术的发展在学术和工业界都取得了飞速的发展。

在 GMIC 2020 阿里文娱技术专场，阿里文娱高级算法专家若仁，分享了视频搜索技术简介、多模态在视频搜索的应用，希望对相关的算法同学能带来启发。

考虑到大家来自不同的业务领域和技术方向，我会先简单介绍优酷视频搜索的业务背景，同时快速介绍搜索的基本评估指标，以及搜索系统的算法框架，还有相关性和排序模型，让大家有一个更全面的认识，后面会重点介绍多模态视频搜索的相关技术。

一、阿里文娱搜索业务是什么?

搜索团队为整个阿里文娱提供一站式的搜索服务，服务范围包括优酷 Phone 和 OTT 端，还包括大麦、淘票票。涉及的检索内容，从影剧综漫的长视频影视库，到覆盖社会各领域的 UPGC 视频。此外，影人和演出场馆的也在搜索服务范围。以优酷为例，我们有数亿的视频资源，不仅包括平台购买版权的 OGC 视频，更多是用户上传的 UPGC 视频。 视频的存储、计算以及分发，比文字更具挑战。



搜索技术的用户价值主要体现在两个维度：

一是工具属性。就是用户将搜索服务作为寻找内容的工具，目标是“找准，找全”，即“搜的到，搜的准”。从这个维度去评估搜索效果的好坏，需要一系列的体验类指标，比如跳出率、相关性，以及时效性和多样性。这些都是搜索通用的技术指标；所谓可播性指在应用上能播放，这是全网视频搜索特有的，受内容版权和内容监管多方面的原因，有一些内容是平台无法播放的。 此外，我们还会定期做人工评测，做横向和纵向比较。

二是分发属性。让用户消费更多的视频内容，有更多 VV(观看视频数) 以及 TS(消费时长) 的引导。这些指标对于垂直搜索，是非常重要的，也是用户满意度最直接的衡量。对于平台来说，搜索还能支持平台的宣发和商业价值，实现广告 / 会员的商业价值，前提是将用户体验做好。

搜索算法框架，由下到上依次是数据层、技术层、内容召回、多媒体相关性、排序、意图。

1）数据层：最基础是视频内容的数据，我们从视频内容中抽取出对应的知识，包括实体、实体之间的关系以及属性。通过内容组织的方式，以图谱知识去指导我们做聚合，从时效性的维度做聚合，从多种维度将内容组织起来。

2）技术层：在数据基础之上，利用 CV 和 NLP 技术，支撑上层内容召回和相关性，排序，以及对 Query 的意图理解。

3）召回层：对多媒体内容理解是难点，下文会详细展开讲。

4）相关性：包括基础相关性 / 语义匹配技术。

5）排序层：按照体验和分发等维度，去提升搜索整体体验。排序利用机器学习排序学习的方式，去提升分发效果，此外还要优化体验类目标，如时效性，多样性等，同时也要实验平台的宣发等目标，是典型的多目标优化场景。

6）意图：对 Query 意图理解，首先要对 Query 做成分分析，标明 Query 各成分是什么，是节目名还是剧集信息。然后要建立细粒度的意图体系，对用户表达的意图去做深层次的意图理解，从而更精准的指导召回，相关性和排序。

多媒体内容理解是视频搜索的重点，视频内容传递的信息是非常丰富的，是不可能用标题的短短十几个文字描述全面的。用户在检索表达时，需求的差别非常大，这就是天然的语义鸿沟。所以我们不能把视频当作黑盒子，需要利用 NLP 能力、CV 的能力以及其他技术能力对视频内容有全面的分析解构。

二、相关性和排序模型

1. 挑战

视频搜索相对通用搜索是有特殊挑战的。第一个挑战是内容相关性匹配，下图中前两个 case，体现出用户表达的 Query 和视频标题不是那么相关，需要通过对内容理解分析，通过对它的元信息的丰富，建立起内容相关性。

搜索相关性

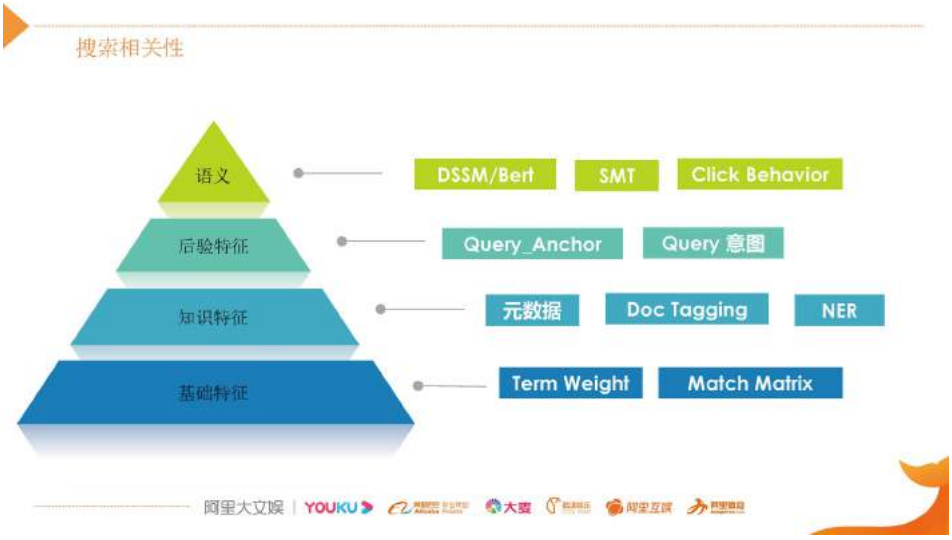
Query	Doc	挑战
佟丽娅主演的电视剧远大前程	发布会回顾 上海滩四大亨登场-赵立新、倪大红、刘奕君、果靖霖	内容理解
变形计2017姚金冬	姚金冬和新"爸爸"对话陷入无尽的尴尬	
法不容情国语	公共用地被占私用，法不容情，拆不容缓！	NER
中国餐馆电视剧	中国餐馆放着抗日神剧	
如何调车后视镜	一分钟学会倒车入库侧方停车看右后视镜科目二反光镜怎么调	语义
怎样去除牙齿污垢	牙黄口臭没自信？学会这一招，轻松去除牙齿上的黄渍污垢，再也不用花钱去洗牙	

阿里大文娱 | YOUKU 大麦 阿里互娱

如“变形计 2017 姚金冬”，视频标题中只有“姚金冬”，实际上通过视频内容的理解，可以知道“姚金冬”和“变形计”，并且是 2017 年的。通过内容理解和 IP 指纹，把 IP 周边视频，如切条或二创视频，和 IP 建立起关联关系，能大大丰富视频的元信息，提升内容相关性匹配度。

第二个挑战是实体知识匹配，我们要借助于视频标题的结构化去理解，用 NER 方式抽取出来，同时也需要 CV 的技术去辅助 NER 识别的准确率。比如“法不容情国语”，QP 端理解出“法不容情”是一个节目名，这就是 Query 的成分分析；用户上传的是“公共用地被占私用，法不容情”的社会问题，需要对文档端去做结构化的理解，要理解出“法不容情”在上下文中不是节目，而是其他意思，然后在相关性匹配的时候，利用这些先验知识做更好的相关性判断，从而决定相关性匹配度。 此外，“中国餐馆电视剧”这个 case 也类似，用户找的是《中国餐馆》节目，而不是需要检索出“中国餐馆放着抗日神剧”。

第三个层面的挑战是语义匹配，当然是通用搜索也有语义匹配问题，一些语义类/how to 类的知识匹配，要去做语义和更全面的分析，比如利用内容理解和实体知识的辅助补充等，才能做好语义匹配。



2. 解法

我们整个视频搜索相关性是从四个维度去做：

1）基础特征：基于基础文本的特征匹配，包括 term weight、匹配矩阵等。

2）知识特征：通过内容理解，以及视频自身所带的元信息，例如视频中的人物，所关联的节目相关的元信息，以及针对视频标题所做的结构，比如我们抽取出哪部分是人物，哪些是 IP 名，哪些是游戏角色等。标题结构化之后，根据 Query 成分的理解，支持在知识层面去做匹配。

3）后验特征：因为用户去搜索 Query 之后，搜索结果之间会产生交互，形成 Query 和 Doc 交互特征。Query_Anchor 以及通过这些交互特征能够指导 Query 意图的理解，把他们作为这种后验关联的一些特征，能够支持我们这种意图匹配。

4）语义：是文本层面的语义匹配，利用 DSSM 语义模型和 Bert 语义模型，做离线和在线的语义匹配模型。除了这种匹配层面之外，还要支持语义召回。通过 SMT 和点击行为分析等技术，进行语义扩展，扩大召回语义内容的范围，利用它们形成的特征更好的做好语义匹配。

希望通过前面两个 slide，能够让大家更好的理解视频搜索相关性的挑战和解法。

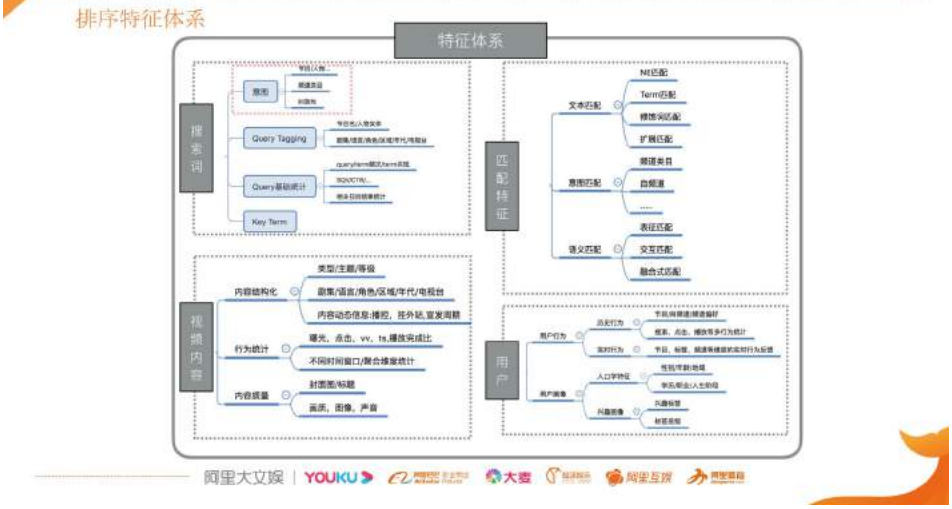
3. 相关性数据集构建和特征体系

全面准确的发现问题是解决问题的基础。相关性数据集的目的是给相关性算法提供 ground truth，标注是重点。相关性标注数据集的标注规范较复杂，标注样本量比较大的，通过外包进行人工的标注，重点需要关注的是标注质量和标注成本。根据标注规范不仅要去标注样本的等级，对同等级下样本的还需要标注偏序关系，质量的把控特别关键。对于成本来说，需要有高效的样本挖掘机制和方法。

如图中左侧所示，可以通过 Active Learning 的思想来加快标注的效率提升，同时也能提升标注质量。 我们可以基于训练集不断地迭代线上模型，对于线上模型预测不是那么准确的样本，可以提供给外包同学去检测标注，形成一个快速的迭代闭环，提升训练的精度。大家用这种方法去做，能够大大提升整个标注的质量和效率。

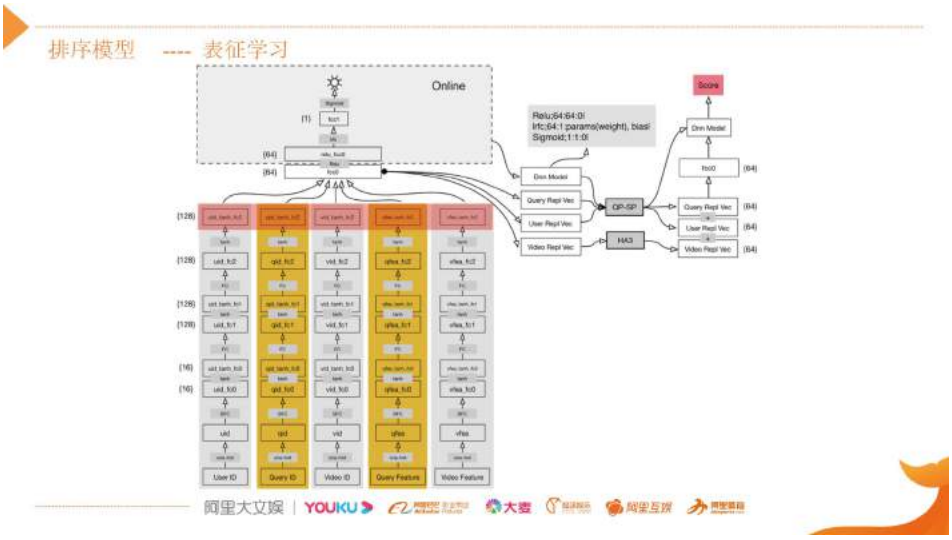
右侧是相关性的技术特征，最上层是 Query 相关的特征，有常规类的，非常基础的文本特征；有意图理解输出的特征，比如说人物、节目、宽泛、How to 类意图等；文档端的先验特征包括从文档标题解构出来的特征以及基础文本特征，除了先验特征还会使用前面说的后验特征；最下层是 Query 和 Doc 的匹配类特征，匹配特征也会分基础匹配、意图匹配、语义匹配、知识匹配这几个维度。

4. 排序特征体系



1. 搜索词特征组：搜索词以及匹配特征这些特征类别，是搜索领域通用的。
2. 匹配特征组：有一些特征是平台特有的，比如视频的实时播控、内容宣发特征。
3. 视频内容特征组：内容质量对于我们的平台非常重要，因为每天上传视频量非常大，需要做好内容质量的评估，才能更好地指导冷启动的分发。我们人工智能部有一个 CV 团队，负责为我们提供高质量的特征，从封面图、标题、画质 / 图像 / 声音各模态去评估视频质量。
4. 用户特征组：用户行为特征，用户画像及用户行为的表征学习特征主要用在一些宽泛搜索场景。例如频道页的搜索排序、OTT 宽泛意图排序等。

接下来分享 2017 年，我们和达摩院在搜索上落地的表征学习排序方案。



第一层是对特征域编码层，按照用户、搜索意图、视频三元素。在用户维度，划分了用户 id 域、用户观看视频序列域；搜索意图维度，划分了搜索词 id 域、搜索词视频表达域、文本编码域。视频维度，划分了视频统计特征域、视频文本编码域、视频 i2i 域。

第二层和第三层不同特征域间网络结构相互独立，通过稀疏编码优化的全连接层对第一层的高维特征域进行降维，把高维信息投影至低维的向量空间中。

通过第三层全连接层对域内信息的二次编码，输出域内特征向量。

通过第四层把 concat 层链接起来，对域间的 id 特征向量、行为特征向量、文本特征向量和观看序列特征向量做多模态的特征向量融合。

之后经过两层的全连接网络实现对给定用户和搜索意图下每个视频的排序分值的预测。这个模型是内容分发的一个排序模型，它同时还会结合相关性模型，时效性，以及视频质量等从多维度模型融合，来决定最后的排序。

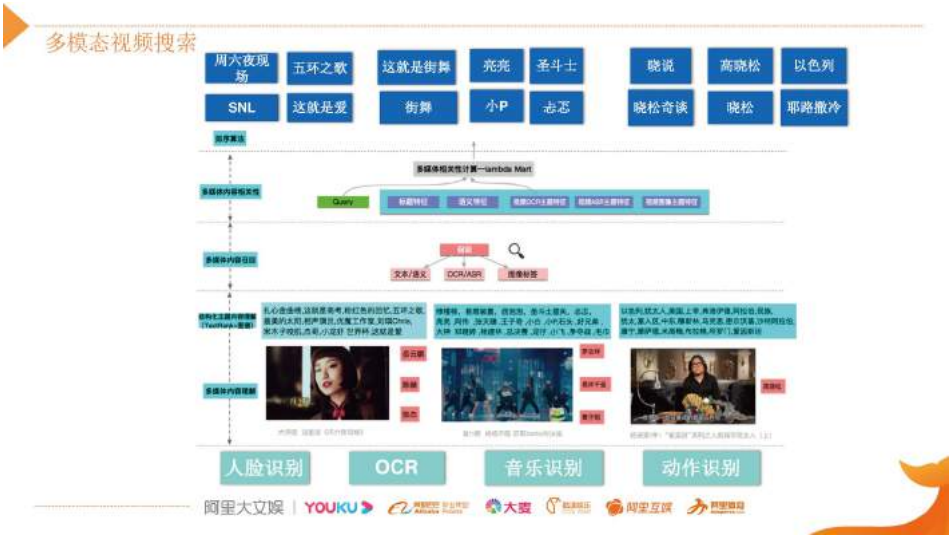
三、多模态视频搜索

基于标题和描述等文本信息的检索，会遇到很多困难。

首先是单模态信息缺失，用户在上传 UGC 视频的时候，标题是比较简单，很难将丰富的视频内容表达清楚，有时这些文字信息还和视频内容是没有关联关系的；

其次是用户搜索意图越来越多元化，即使是版权视频的搜索也不再集中于节目名字的搜索，社交与互动的需求逐渐增长越来越多；

其次还有 TO B 侧的需求，也就是内容二创的用户，需要去找各种各样的视频片段，视频素材，这部分的需求也是需要用多模态技术去支撑。



基于多模态技术的搜索，将语言、语音、文字、图像的各种模块集中起来，综合这些信息来理解，能够方便用户更好地找到所需内容，得到更好的搜索体验。目前在工业界和学术界对多模态搜索的研究热度都非常高，学术界有很多方案是基于将 Query 和视频映射到一个中间 Concept 的空间，然后在中间维度做相似度的匹配排序。此外，也有 VQA/GQA 等各类基于视频问答的数据集去推动自然语言和图像的关系推理，最近 Video/Visual Bert 的端到端的解决方案也有了很大的突破。

学术界喜欢这种端到端的解决方案的，对于短小视频的理解，可能是个不错的方案。但针对长视频，这些方案是很难真正做到准确的理解的。因为长视频包含内容信息是更多维，更宽泛。此外在工业界搜索引擎是需要有可解释和可控性的，很少是采用单一端到端的解决方案。所以我们优酷的多模态搜索采用的技术方案是：

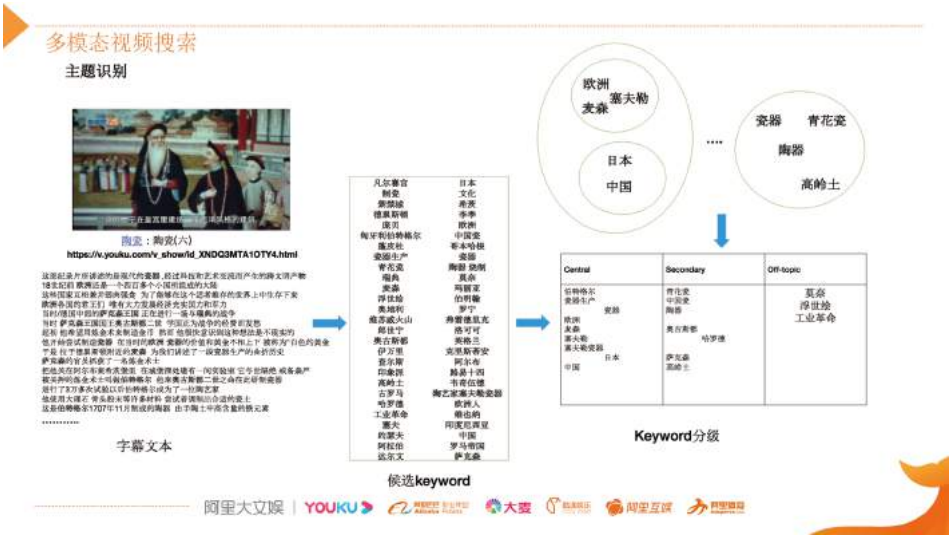
1. 利用 CV 算法技术，将其他模块的信息降维到文本模态。

2. 通过我们多模态内容检索的技术实现召回。

3. 再从上层通过内容相关性和排序技术，达到用户对内容各维度检索需求的满足。

例如，基于人脸识别的技术，将视频中出现的明星人物，如《这就是街舞》视频中识别出易烊千玺、黄子韬等；此外通过 OCR/ASR 的技术，识别各视频中的对话内容转化成文本。然后基于文本去做结构化理解。

结构化的文本，需要有系统性的理解和组织，用后面会讲到的关键词抽取技术，把它理解好并形成我们的一些内容主题；同时还会利用音乐识别、动作识别、以及场景识别，情绪识别等 CV 技术，不断丰富解构内容，实现用户在上面去各种组合搜索的时候，我们都能够召回，还能排的比较好。



这个 slide 主要是用一个案例来介绍我们在多模态视频搜索时，内容关键词是怎么更好的组织，视频内容降维成文本之后，怎么能够去做好这些文本内容的组织理解。

首先，从案例看到，内容关键词的词库是非常非常大的，此外内容和关键词属于多对多的关系。我们要通过各种关键词的抽取技术抽取候选的内容关键词，并且要扩大候选词来源的一个多样性，比如基于 "NER" 的方法能确保抽取的内容关键词是百科类实体名称，有较广泛的知识内涵；“新词发现”方法会综合 Ngram 以及语言模型 (LM) 等多种基础能力扩大对未知知识领域的挖掘。

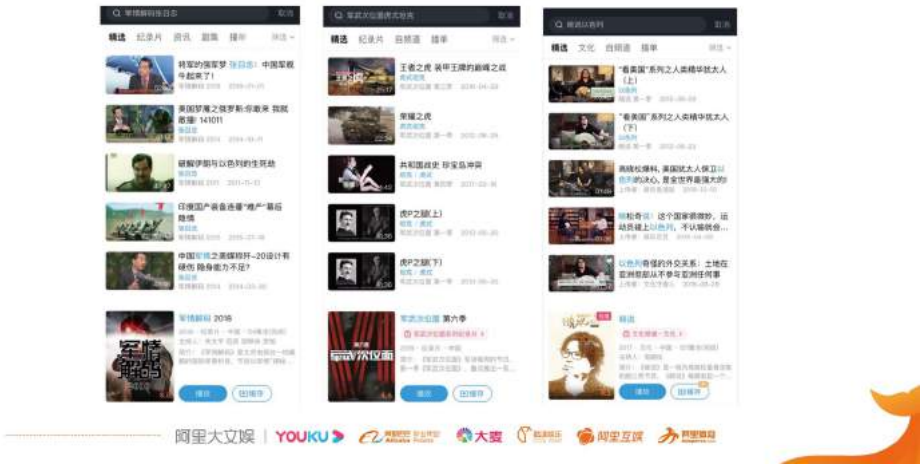
候选关键词是一个不断扩充的过程，随着我们在视频内容理解的维度扩大，候选关键词的来源会越来越丰富。在丰富的内容候选关键词基础上，根据内容候选关键词和视频内容相关程度构建分类模型预测不同的等级，最相关的是核心内容关键词，其次是相关内容关键词以及提及内容关键词，然后整个在关键词分级的核心特征是除了文本特征之外，还会采用音频 / 视频表征网络生成的一些多模态特征来共同训练，来提升预测关键词相关度的准确率，把关键词和内容表达的关联度预测更精准。

在过程中会看到这么做存在的一些问题，以图中视频为例，该视频主要是讲欧洲瓷器的发展史，但是该视频文本标题是“陶瓷：陶瓷 (六)”，非常简短的描述，对它做内容理解降维成文本后，能够利用上面讲到的技术抽取内容关键词“塞夫勒”，“麦森”，但是如何能够把“塞夫勒”、“麦森”和“欧洲”关联起来，知道这个视频讲的是欧洲瓷器发展史，而不是中国或者日本；此外对于瓷器领域知识实体，“陶器”，“青花瓷”，“高岭土”，怎么把它们和“瓷器”概念关联起来。

这些都是需要有知识图谱 (KG) 知识支撑的，这就需要 KG 实体知识库涵盖广泛的领域，需要有全行业的丰富实体，才能帮助提取核心内容主题。另外像抽取的内容关键词“伯特格尔”是个人名，但是要用什么技术能把它和内容主题相关程度识别准确，知识库不一定能收录，单纯通过频次也不一定能理解准确，但是“伯特格尔”被“他”指代提及多次，需要有这种指代推理能力，把这样的关系理解出来。有了这些关系的理解，才能基于内容关键词之上去理解整个的内容事件、内容主题、以及内容

故事线等不同层级的抽象，才能够更全面的理解视频，然后来更好的去支撑上层的召回匹配和排序。

效果案例



目前，我们做的这些探索都上线了，在线上能看到效果。像用户搜索“军情解码张召忠”，排前面的这几个视频内容都是“张召忠”主讲的，但是在标题文本里面是没有的，是通过内容理解的方式能够把它抽取出来；像“军武次位面虎式坦克”，“虎式坦克”是用户是要找的，但是在视频标题中都是“荣耀之虎”，“虎 P 之腿”，这些视频里面是针对“虎式坦克”有详细的内容介绍，通过内容理解能够将用户的需求和内容关联起来，能做比较好的召回和排序；最右边的是高晓松老师的“晓说以色列”也是这类，大家可以在优酷 APP 上多做一些体验。

加入交流群

阿里巴巴文娱技术
钉钉交流群

关注我们

阿里巴巴文娱技术公众号