

به نام خدا

# گزارش پروژه درس بینایی ماشین سه بعدی

گردآورندگان

هیراد داوری

سیدعلی حسینی

۱۴۰۳ بهار

## فهرست مطالب

2.....	فهرست مطالب
3.....	مقدمه
3.....	چکیده
3.....	مدل‌ها
4.....	Florance 2
4.....	DepthAnything-v2
5.....	YOLOv8
5.....	روش ارائه شده
7.....	مقایسه و نتیجه گیری
7.....	ضمیمه
8.....	DepthAnything
8.....	Florence 2 (prompt: A suspended load)
9.....	YOLOv8 (persons)
9.....	نتیجه نهایی

## مقدمه

در صنعت، حمل و نقل بارهای سنگین توسط جرثقیل‌ها از اهمیت ویژه‌ای برخوردار است و عدم رعایت اصول ایمنی می‌تواند به بروز حوادث جبران‌ناپذیر منجر شود. یکی از ریسک‌های مهم در این زمینه، قرار گرفتن افراد در زیر باری است که توسط جرثقیل در حال جابجایی است. این وضعیت می‌تواند باعث آسیب دیدگی‌های جدی و حتی مرگ و میر شود. به همین دلیل، شناسایی و جلوگیری از حضور افراد در فضای زیر بار معلق اهمیت بسیاری دارد. کد پروژه در این [لينك](#) موجود است.

## چکیده

کاری که در این پروژه انجام شد استفاده از چند CNN پردازش تصویر و استفاده از آنها در یک pipeline همراه برنامه‌های خاص منظوره بود تا بتوانیم فرآیند تشخیص ناحیه خطر زیر بار آویزان (suspended load) را مشخص کرده و ببینیم که آیا انسانی در این ناحیه قرار دارد یا خیر. درواقع این مدل‌ها برای تشخیش اجسام از این دست و برای پردازش فضای سه بعدی train نشده اند اما با قرارگیری آنها در این pipeline توانیستم این کاربری را از آنها استخراج کنیم

## مدل‌ها

از مدل‌های زیر برای پیاده سازی این پروژه استفاده شده:

- Florence 2
- DepthAnything v2
- YOLOv8

## Florance 2

مدل 2 Florence به توسعه شرکت مایکروسافت، نگرشی جدید به هوش مصنوعی در پردازش تصویر است. این مدل داخل خود یک پردازشگر زبان طبیعی (NLP) دارد که بخشی از آن مربوط به تولید embedding برای توکن های استفاده شده در جملات است. ایده Florence این است که برای تصاویر با رویکرد attention-based embedding های خاص و محلی ایجاد کرد و سپس این embedding ها با embedding های بسته آمده از کلمات مقایسه کرد. درواقع به این صورت مدل Florence می‌تواند تصاویر را با کلمات توصیف کند! یکی از کاربردهای Florence ایجاد caption برای تصاویر است که می‌تواند بسته به نیاز جزئی یا کلی باشد. وقتی این کار انجام شد می‌توانیم با گرفتن یک توصیف از کاربر (prompt)، بخش هایی از تصویر که به توصیف مربوطند را پیدا کنیم. مزیت بزرگ Florence این است که اندازه کمی دارد اما نسبت به این اندازه، توانایی آن در segmentation تصاویر و object detection بسیار بالاست به خصوص با توجه به این که می‌تواند با توجه به prompt که از کاربر دریافت می‌کند خروجی ایجاد کند.

## DepthAnything-v2

مدل DepthAnything مدل توسعه یافته توسط شرکت TikTok است که به نوبه خودش یک جهش فنی محسوب می‌شود. این مدل در حال حاضر یکی از بهترین مدل‌های کاربردی برای پردازش عمق در تصاویر است. نکته مثبت این مدل دقت بالای آن بدون توجه به کاربرد و نوع عکس است.

## YOLOv8

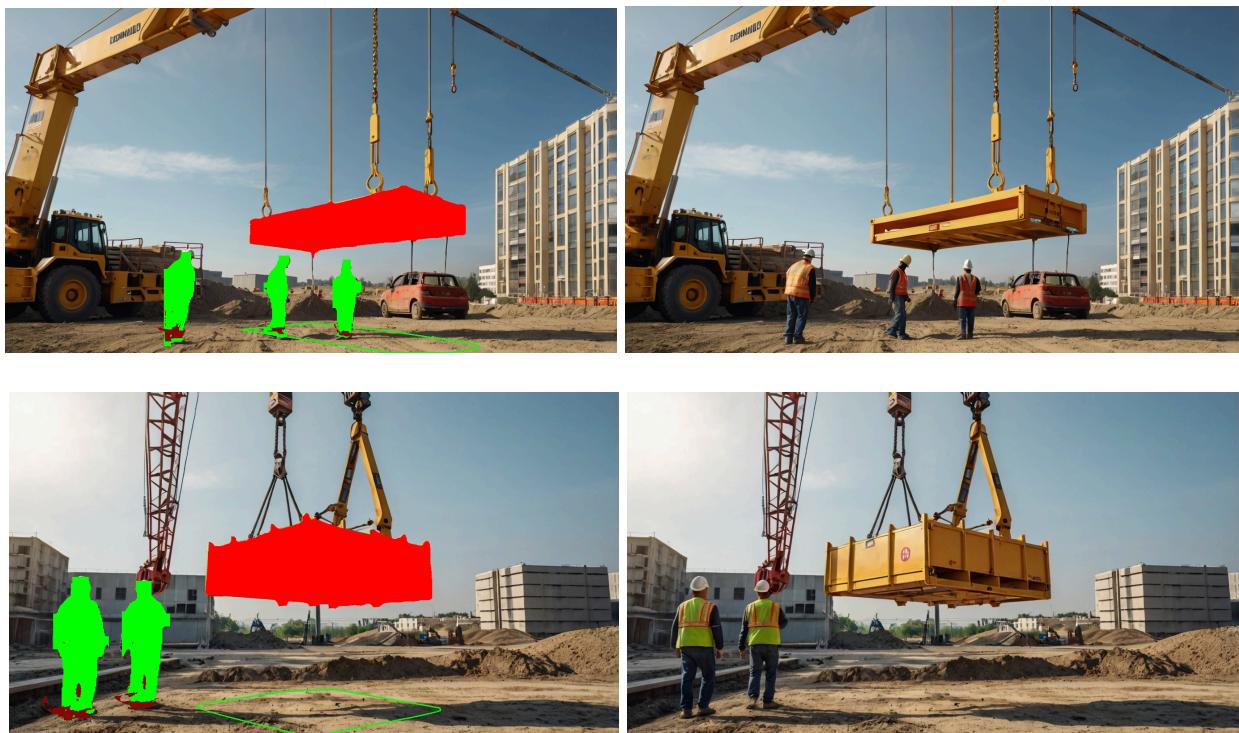
مدل YOLO یکی از مدل های معروف و سریع object detection است. دارای یک CNN عمیق است که با ابعاد مختلف عرضه می شوند. این مدل از قبل روی دیتاست COCO آموزش دیده و به همین خاطر دامنه وسیعی از label ها را پشتیبانی می کند. مزیت اصلی آن در کاربری تشخیص شی، مثلاً نسبت به florence سرعت بالای آن است. تأخیر زمانی این مدل به حدی پایین است که از این مدل در پردازش video هم استفاده می شود، اما مدل florence بعضاً تأخیر در حدود چند ثانیه دارد.

## روش ارائه شده

در این پژوهه تصاویری از صحنه های ساخت و ساز داریم. به طور کلی فرض کنید که این تصاویر frame هایی از یک دوربین فیلمبرداری هستند. ابتدا تصویر به مدل DepthAnyting داده می شود تا برای هر پیکسل تصویر عمق تخمینی آن بدست آید. سپس از این عمق ها استفاده می کنیم تا نقاط را در فضای سه بعدی بازسازی کنیم. البته که متريک فضای ایجاد شده با واقعیت همخوانی ندارد اما این موضوع در کاربرد ما اهمیت چندانی ندارد چون پس از انجام محاسبات در این فضای سه بعدی نتایج دوباره به افکننده می شوند. سپس تصویر به مدل Florence داده می شود. همراه با تصویر باید یک Image plane prompt با معنی به آن بدھیم تا پیکسل های مرتبط به suspended load را پیدا کند. پس از این مرحله باید پیکسل های مربوط به زمین را پیدا کنیم. برای این کار از تعدادی نقاط fix شده و یافتن پاسخ LSE برای فیت کردن یک صفحه به نقاط استفاده کردیم. در انتها مدل YOLO وظیفه پیدا کردن پیکسل های مربوط به افراد داخل تصویر را دارد. از مدل Florence به دو علت استفاده نشد، یک این که پس از آزمایش با این مدل به این نتیجه رسیدیم که این مدل افراد را با اطمینان خوب پیدا نمی کند و polygon هایی که ارائه می دهد معمولاً دارای redundancy و فضاهای خالی اند. بعض اصلاً افراد را پیدا نمی کند و دوم این که این مدل به علت سرعت پایین ترش نسبت به دو مدل دیگر bottleneck این پژوهه

محسوب می شود و YOLO این کار را سریع تر انجام می دهد، به طور کلی اجسام آویزان (عموماً از جرثقیل) با سرعت کمتری نسبت به افراد حرکت می کنند که یعنی می توان به ازای چند فریم از بررسی حرکت افراد، یک بار حرکت جسم را بررسی کنیم. برای مشخص کردن danger zone باید نقاطی که در polygon مربوط به suspended load است را روی صفحه به دست آمده project کنیم و پس از یافتن convex hull آنها، این شکل را به image plane بیفکنیم.

در شکل زیر، چند مثال از اجرای الگوریتم را مشاهده می کنید. نقاط زرشکی، سایه های عمودی افراد روی زمین است و چند ضلعی سبز سایه های عمودی suspended load بر روی زمین است. همان طور که می بینید، در سناریوی اول، دو نفر زیر suspended load قرار دارند که به درستی تشخیص داده شده اند. در تصویر دوم (پایین) هم، افراد زیر suspended load نیستند و این بار هم این قضیه به درستی تشخیص داده شده است.



## مقایسه و نتیجه گیری

مسئله پیدا کردن falling object danger zone یک نمونه از بسیار مسائل مهمی است که پیشرفت در شاخه هوش مصنوعی می‌تواند در آنها کمک کند. شاخه هوش مصنوعی به پیشرفت‌های چشمگیری در زمینه‌هایی که عموماً جنبه کلی دارند رسیده، اما به بسیاری از مسائل خاص منظوره رسیدگی نشده. در مواجهه با این مسائل راه حل‌های هوش مصنوعی معمولاً شامل تهیه تیم و یافتن راه حل جدید برای این موضوع است اما همینطور که در این پژوهه دیده شد، مدل‌های general purpose که در حال حاضر وجود دارند، با تعامل با یک دیگر می‌توانند با هزینه پایین‌تر و با development time کمتر به راه حل‌های specific purpose برای این مسائل تبدیل شوند. در واقع به جای تحقیق و صرف هزینه برای ایجاد یک مدل جدید پیچیده که وظیفه یافتن danger zone را به عهده دارد، می‌توانیم از چند مدل مختلف که برای segmentation استفاده می‌شوند و قرار دادن آنها در یک application، از آنها در scope فراتر از چیزی که برای آن ایجاد شده اند استفاده کنیم.

## ارزیابی مدل

برای ارزیابی مدل، MeanIoU را برای Depth+Florence2 و Depth+YOLO8 محاسبه کردیم:

Person Mean IoU: 0.24

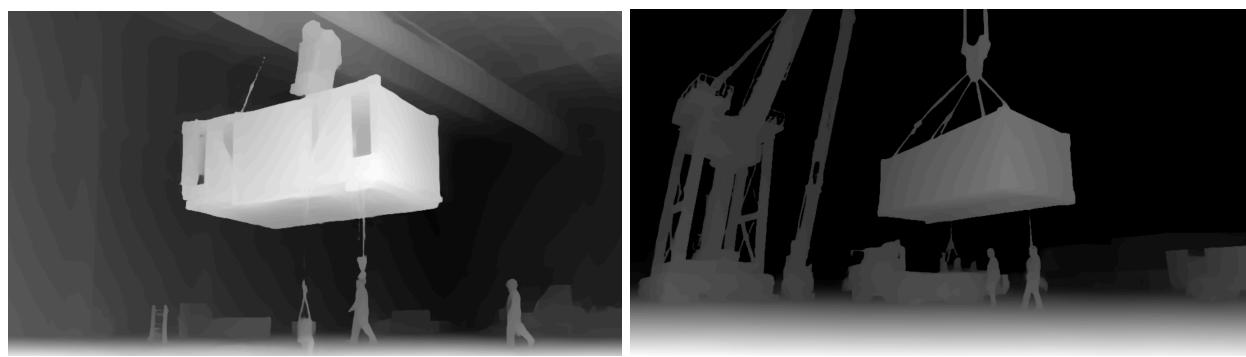
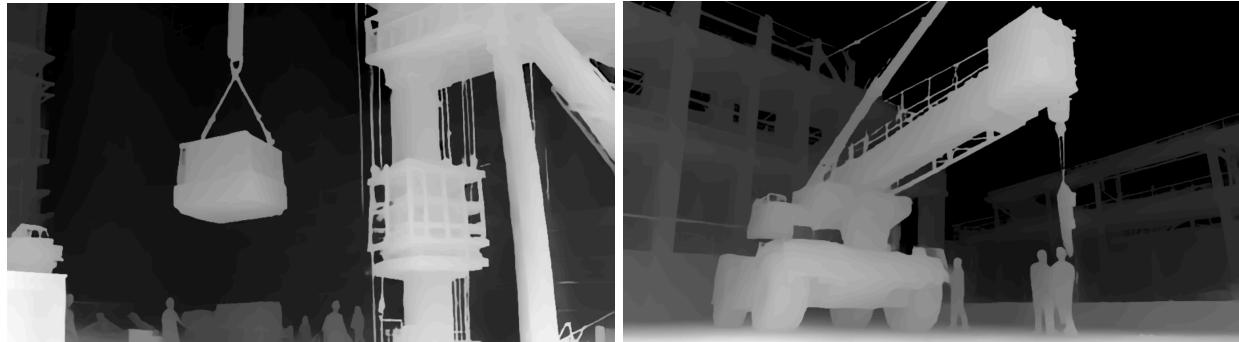
Person Mean IoU: 0.12

گفتنیست که با وجود مقادیر کم این متريک‌ها، دقت مدل در تشخيص افراد زیر suspended load ۱۰۰ درصد است.

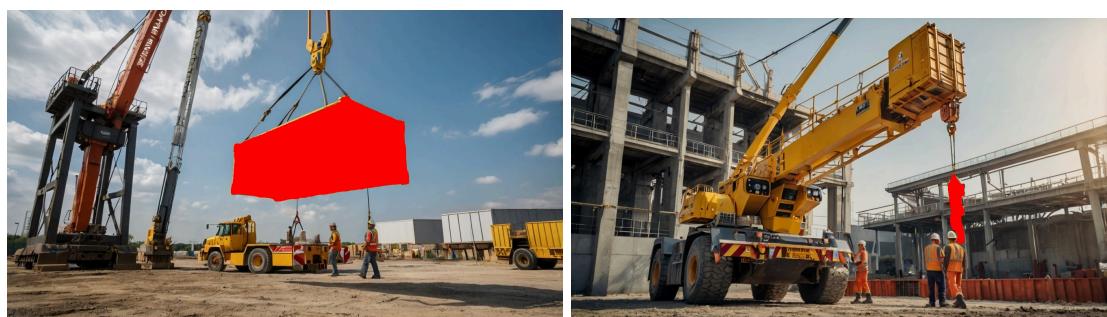
## ضمیمه

در این بخش، تصاویر مربوط به اجرای الگوریتم‌های مختلف آورده شده است:

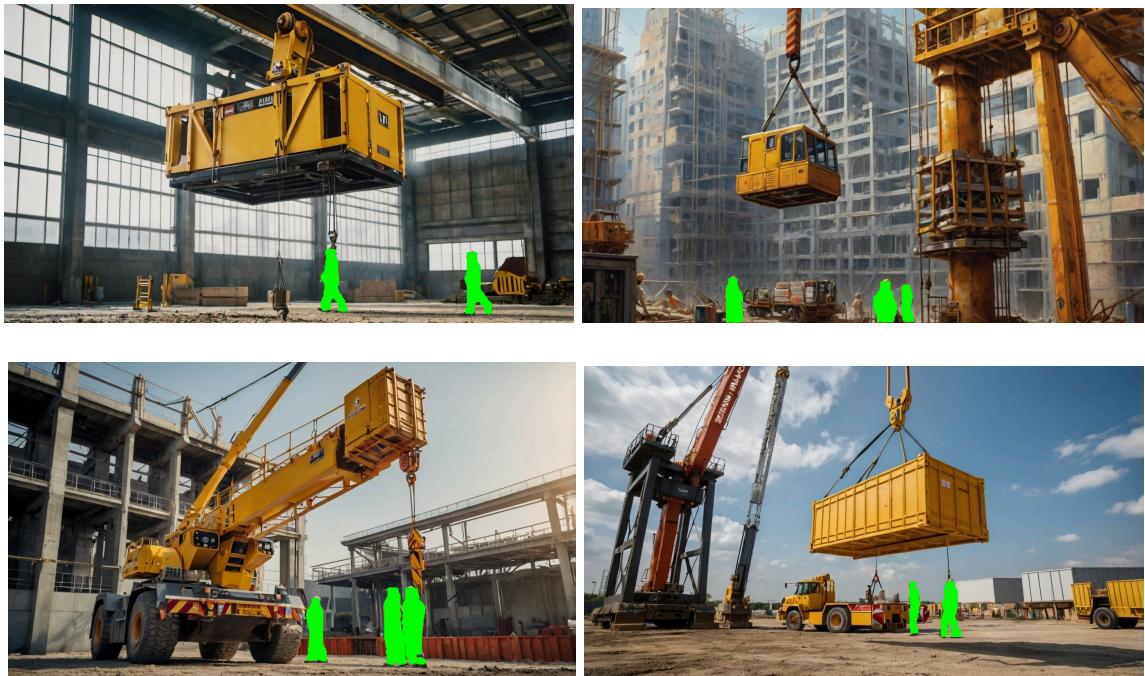
DepthAnything



Florence 2 (prompt: A suspended load)



YOLOv8 (persons)



نتیجه نهایی

