Vision-language tasks and variants

Early VLMs focus on image-level visual recognition tasks, whereas recent VLMs are more general-purpose, which can also work for dense prediction tasks that are complex and require localization related knowledge.