Language as a Camera

Phillip John Isola
Associate Professor, MIT EECS

Abstract:
Visual content can be conveyed in many ways. It can be photographed and captured in an array of pixels, or instead it can be described through text rich in imagery. Computer vision has traditionally only dealt with the former format, leaving language processing as the domain of other fields. In this talk I will reconsider this choice: should computer vision also deal with language as a fundamental visual format? I will share our recent work asking: what do language models know about the visual world? Are they good models of visual data? What kinds of visual structures do they represent? And how can they be leveraged to improve vision systems.