

# Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and  
James M. Rehg

School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA  
Feunji chong, nataniel . ruiz, ywang751, yzhang467, agata, rehg@gatech.edu

**Abstract.** This paper addresses the challenging problem of estimating the general visual attention of people in images. Our proposed method is designed to work across multiple naturalistic social scenarios and provides a full picture of the subject’s attention and gaze. In contrast, earlier works on gaze and attention estimation have focused on constrained problems in more specific contexts. In particular, our model explicitly represents the gaze direction and handles out-of-frame gaze targets. We leverage three different datasets using a multi-task learning approach. We evaluate our method on widely used benchmarks for single-tasks such as gaze angle estimation and attention-within-an-image, as well as on the new challenging task of generalized visual attention prediction. In addition, we have created extended annotations for the MMDB and GazeFollow datasets which are used in our experiments, which we will publicly release.

**Keywords:** Visual attention · Gaze estimation · Saliency

## 1 Introduction

As humans, we are exquisitely sensitive to the gaze of others. We can rapidly infer if another person is making eye contact, follow their gaze to identify their gaze target, categorize quick glances to objects, and even identify when someone is not paying attention [19]. Automatically detecting and quantifying these types of visual attention from images and video remains a complex, open challenge. Although gaze estimation has long been an active area of research, most work has focused on relatively constrained versions of the problem in specific predetermined contexts. For example, [31, 18] predict the gaze target *given* that the person is looking at a point on a smartphone screen, [23] predicts fixation on an object *given* that the person is looking at salient object within the frame, [7, 30] predict eye contact *given* that the camera is located near the subject’s eyes, and [24] predicts the focus of a person’s gaze across views in commercial movies which include camera views that follow the actor’s attention. It remains a significant challenge to design a system that can model the visual attention of































