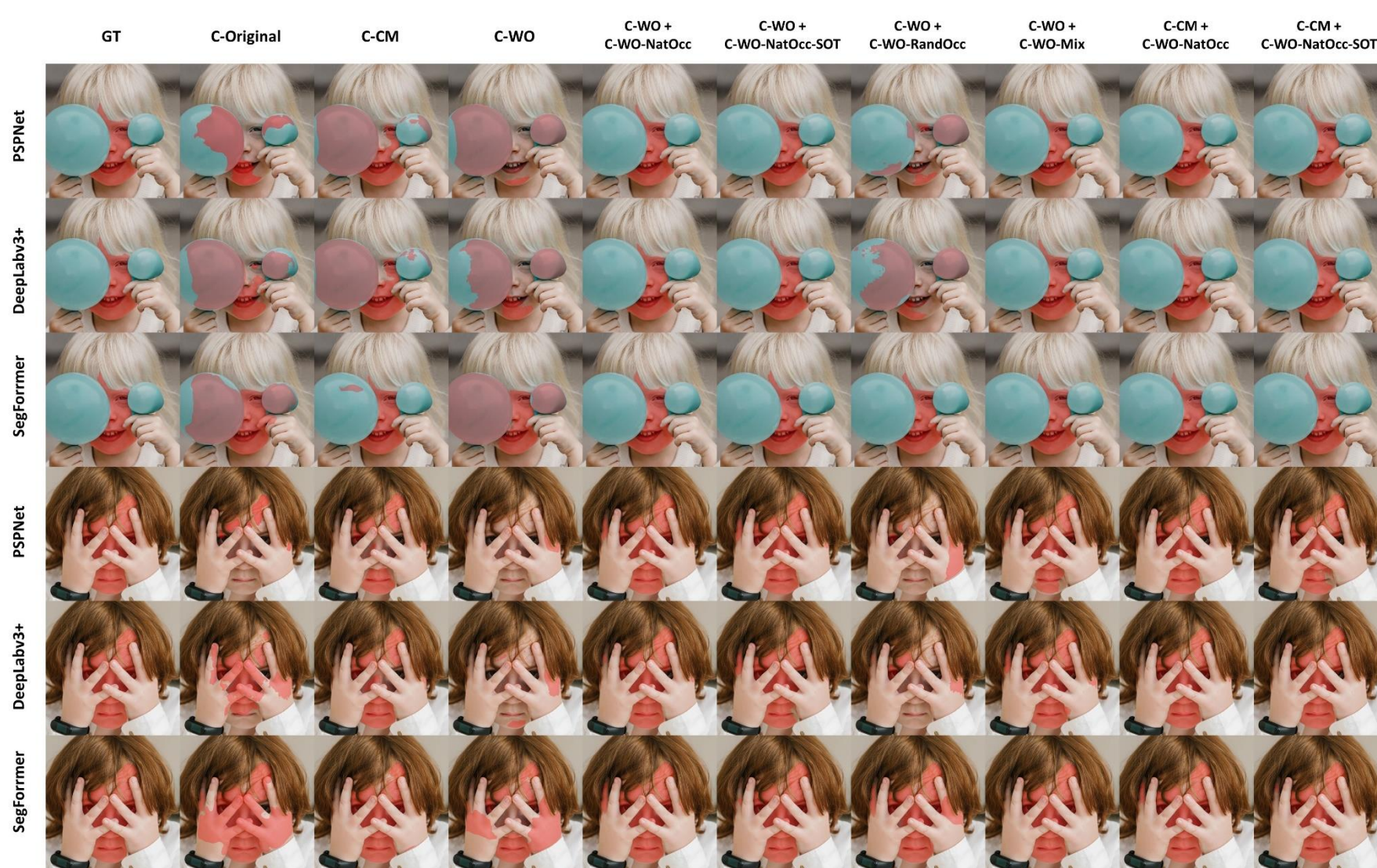# Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets

Student: Kenny Voo Tze Rung   Supervisor: Dr Loy Chen Change

Class	Definition
C-Original	CelebAMask-HQ-WO (Train) and CelebAMask-HQ-O with original masks.
C-CM	CelebAMask-HQ-WO (Train) and CelebAMask-HQ-O with corrected masks.
C-WO	CelebAMask-HQ-WO (Train).
C-WO-NatOcc	One set of hand-occluded (without color transfer) face dataset and one set of COCO-occluded face dataset generated by NatOcc with C-WO.
C-WO-NatOcc-SOT	One set of hand-occluded (with color transfer) face dataset and one set of COCO-occluded face dataset generated by NatOcc with C-WO.
C-WO-RandOcc	Two sets of occluded face dataset generated by RandOcc with C-WO.
C-WO-Mix	Half set of C-WO-RandOcc and one set of C-WO-NatOcc.

Abstract:

This paper performs comprehensive analysis on datasets for occlusion-aware face segmentation, a task that is crucial for many downstream applications. The collection and annotation of such datasets are time-consuming and labor-intensive. Although some efforts have been made in synthetic data generation, the naturalistic aspect of data remains less explored. In our study, we propose two occlusion generation techniques, Naturalistic Occlusion Generation (NatOcc), for producing high-quality naturalistic synthetic occluded faces; and Random Occlusion Generation (RandOcc), a more general synthetic occluded data generation method. We empirically show the effectiveness and robustness of both methods, even for unseen occlusions. To facilitate model evaluation, we present two high-resolution real-world occluded face datasets with fine-grained annotations, RealOcc and RealOcc-Wild, featuring both careful alignment preprocessing and an in-the-wild setting for robustness test. We further conduct a comprehensive analysis on a newly introduced segmentation benchmark, offering insights for future exploration.