Multi Target Mask Encoder Architecture

From Local Embedding Sub-Network

$D_{sub\text{-}network}$

Embedding Layers

Target mask m^t

$(256,256,11)$

$E_{Left\ Eye}$

$E_{Right\ Eye}$

E_{Mouth}

E_{Skin}

E_{Hair} $(64,64,256)$

Concatenated masks and
Facial embeddings

G_m

$(256,64,64)$

To Background Fusing Sub-Network