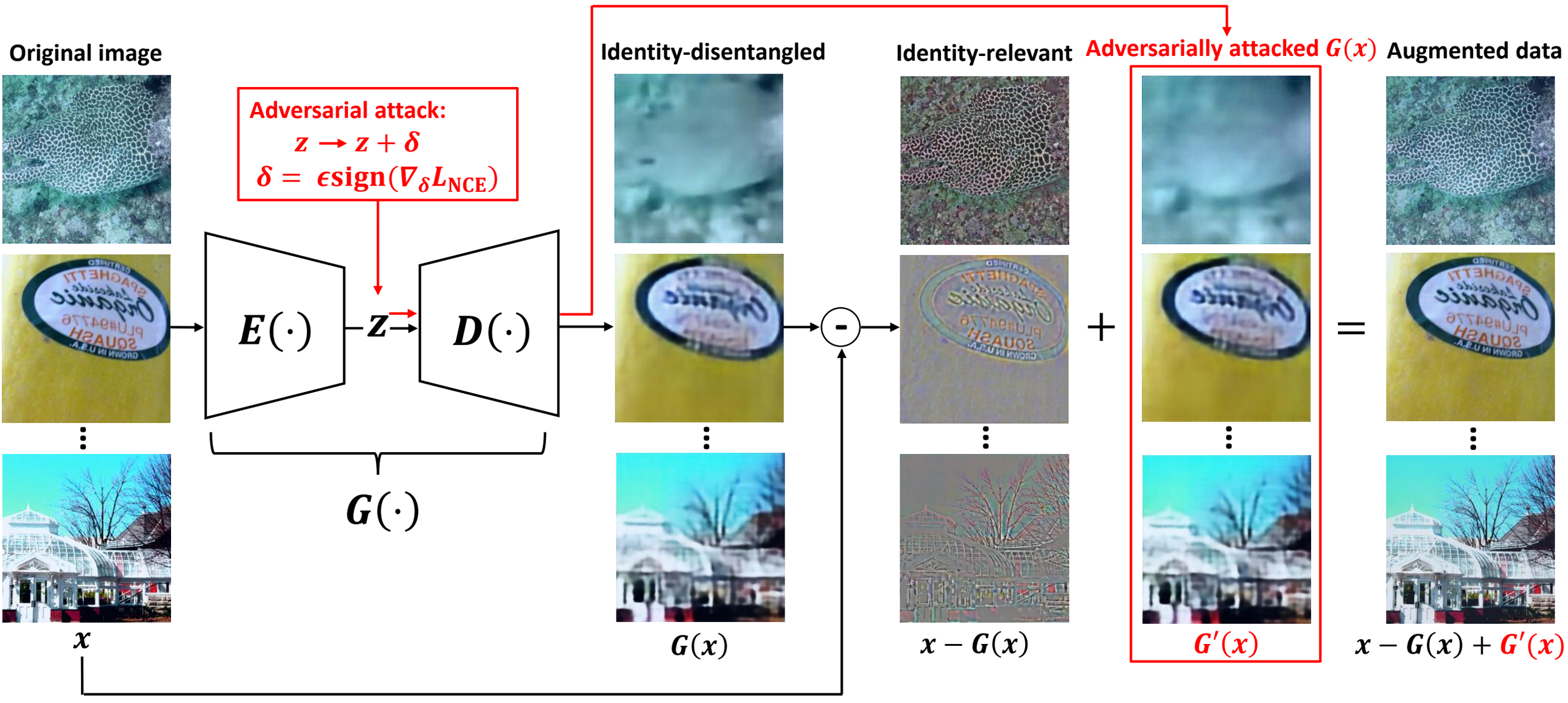
Original image
Adversarial attack:
$z \rightarrow z + \delta$
$\delta = \epsilon \text{sign}(\nabla_\delta L_{\text{NCE}})$
$E(\cdot)$
z
$D(\cdot)$
$G(\cdot)$
Identity-disentangled
Identity-relevant
Adversarially attacked $G(x)$
Augmented data
-
+
=
x
$G(x)$
$x - G(x)$
$G'(x)$
$x - G(x) + G'(x)$