Many labeled/unlabeled corresponding/non-corresponding multimodal data is available, e.g. unlabeled RGB-Depth data.

We formulate RGB-based hand pose estimation as a multi-modal learning, cross-modal inference problem and propose to learn a joint latent representation that leverages other modalities as weak labels to improve RGB-based hand pose estimation.

### Algorithm

**Require:** $x, y, w_1, T$

**Ensure:** $\phi_x, \phi_w, \theta_y, \theta_w$

1. Initialize $\phi_x, \phi_w, \theta_y, \theta_w$
2. for $t = 1, \ldots, T$ epochs do
3. Encode $x$ to $q_{\phi_x}(z_x|x)$
4. Encode $w_1$ to $q_{\phi_w}(z_w|w_1)$
5. Construct $z_{j\text{on}t} = \text{GProd}(z_x, z_w)$
6. Decode $z_x, z_w, z_{j\text{on}t}$ to $p_{\theta_y}(y|\cdot), p_{\theta_w}(w_1|\cdot)$ respectively
7. Update $\phi_x, \phi_w, \theta_y, \theta_w$ via gradient ascent of joint objective
8. end for