Aligning Latent Spaces for 3D Hand Pose Estimation

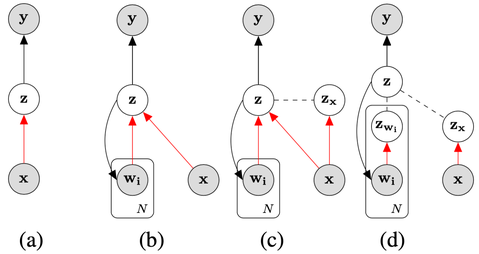
Linlin Yang*[1], Shile Li*[2](*Equal contribution), Dongheui Lee[2,3], Angela Yao[4]
[1]University of Bonn, [2]Technical University of Munich, [3]German Aerospace Center, [4]National University of Singapore

ICCV 2019 Seoul, Korea

Motivation

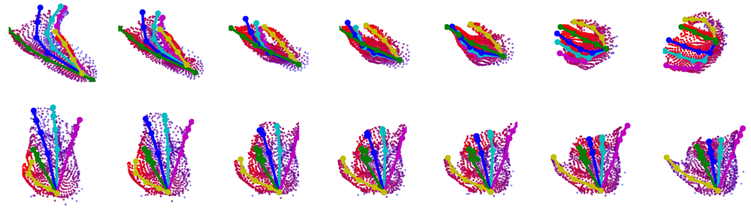
- Many labeled/unlabeled corresponding/non-corresponding multimodal data is available, e.g. unlabeled RGB-Depth data.
- We formulate RGB-based hand pose estimation as a multi-modal learning, cross-modal inference problem and propose to learn a joint latent representation that leverages other modalities as weak labels to improve RGB-based hand pose estimation.

Latent Space Alignment

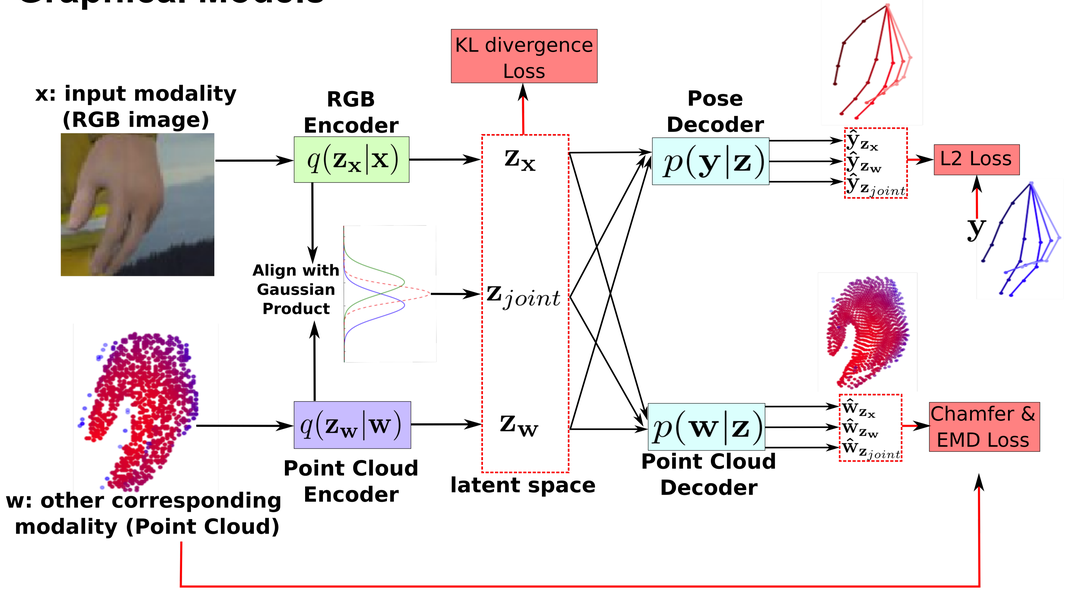


(a) Cross modal.
(b) Extended cross modal.
(c) Latent alignment with a KL divergence loss.
(d) Latent alignment with the product of Gaussian experts.

Latent Space Interpolation



Graphical Models



Algorithm

Require: x, y, w_1,T
Ensure: $\phi_x, \phi_{w_1}, \theta_y, \theta_{w_1}$
1: Initialize $\phi_x, \phi_{w_1}, \theta_y, \theta_{w_1}$
2: **for** t = 1,…,T epochs **do**
3: Encode x to $q_{\phi_x}(z_x|x)$
4: Encode w_1 to $q_{\phi_{w_1}}(z_{w_1}|w_1)$
5: Construct $z_{joint} = GProd(z_x, z_{w_1})$
6: Decode z_x, z_{w_1}, z_{joint} to $p_{\theta_y}(y|\cdot), p_{\theta_{w_1}}(w_1|\cdot)$ respectively
7: Update $\phi_x, \phi_{w_1}, \theta_y, \theta_{w_1}$ via gradient ascent of joint objective
8: **end for**

Pose Estimation