Aligning Latent Spaces for 3D Hand Pose Estimation

monocular RGB 입력으로 부터 hand pose 를 추정하는 것은 어렵다.

기존의 연구는 depth map 과 같은 다른 modality를 이용할 수 있음에도 불구 하고 오로지 RGB 정보만 사용한 monocular 환경에 대한 것이었다.

이 논문에서는 RGB 를 기반으로 한 hand pose estimation을 개선하기 위해 다른 modality를 weak label로 이용한 joint latent representation 학습을 제안한다.

heat map, depth map 그리고 **point cloud**와 같은 여러가지 modality를 embedding 했다.

특히, hand surface의 point cloud를 encoding 하고 decoding하는 것은 joint latent representation의 품질을 향상 시킨다는 것을 알아 냈다.

Introudction

hand pose estimation의 적용 분야

- human activity analysis
- human computer interaction
- robotics

Depth-based pose estimation은 딥러닝을 이용하면서 매우 정확해 졌다. depth 센서가 일반화 됐음에도 불구하고 고품질 depth map은 실내에서만 캡처할 수 있으므로 사용할 수 있는 환경이 제한된다. 게다가, 기존의 RGB 화면뿐만 아니라 단순한 RGB 카메라는 여전히 depth 카메라와 depth 데이터보다 훨씬 더 흔하다.

이와 같이 정확한 RGB 기반 3D hand pose estimation은 여전히 필요하며, 특히 monocular viewpoint 로 볼 때 더 그러하다.

monocular RGB 입력과 관련된 모호성을 다루기 위해, 이전 작품들은 대량의 training데이터에 의존해 왔다. 3D hand pose와 같은 정확한 ground truth label을 얻기란 매우 어렵기 때문에 데이터 세트 크기가 순전히 증가함에 따라 얻는 이득은 거의 없다.

3D hand joint position 을 정확하게 annotation하는 것은 어려운 작업이며, 종종 human annotator간에 합의가 거의 이루어지지 않는다.RGB 영상을 생성하기 위해 여러 가지 방법이 개발되었지만, 합성 데이터와 실제 데이터 사이에는 여전히 큰 도메인 간 차이가 존재하여 합성 데이터의 효용을 제한 되고 있다.

RGB 데이터에 대한 정확한 ground truth은 수집하기 어렵지만 label이 부착되지 않은 RGB-D hand 데이터는 label 부착된 깊이 맵과 함께 활용할 수 있다.

이 논문에서는 RGB based hand pose estimation을 강화하기 위한 약한 라벨로서 multiple modality를 활용하는 것을 목표로 한다.

이 논문에서 다양한 hand data의 modality(RGB image, depth map, point clouds, 3D pose, heat map 그리고 segmentation mask) 를 고려 했고 cross-modal inference 문제로 RGB based hand pose estimation를 formulate 했다. 특히, **multi-modal VAE**를 제안했다.

이 논문에서는 3D pose, point cloud, heat map 등 세 가지 multiple modality에 대해 서로 다른 목표를 도출하고 관련된 hand latent spaces를 정렬하는 두 가지 방법을 보여준다.

장점

1. 무엇보다도 먼저, 빠르게 수렴하여 잘 구조화된 잠적 공간을 만든다 . 이와 비교하여, multi-modal shared latent space은 multiple modality에서 데이터를 추출할 때 변동하는 경향이 있다.
2. 정렬을 통한 학습 계획은 해당되지 않는 데이터와 weak supervision으로 작업하는 데 더 많은 유연성을 제공한다.

결과적인 latent representation은 monocular RGB 이미지로부터 모두 매우 정확한 hand pose를 추정하고 hand surface의 사실적인 point cloud를 합성 할 수 있다

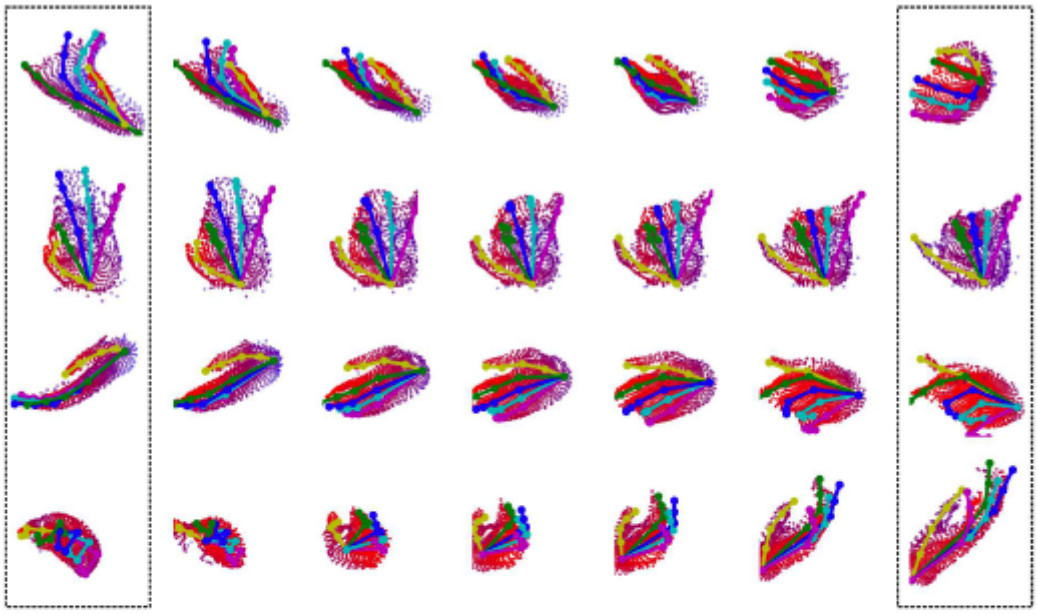


Figure 1: Latent space interpolation. The far left and far right columns (dashed boxes) are generated poses and point clouds from monocular RGB images sampled from the training data. Other columns are generated from linear interpolations on the latent space. The smoothness and consistency imply that different cross-modal latent spaces can be embedded and aligned into one shared latent space.

contribution

- RGB based hand pose estimation을 multi-modal learning, cross modal inference로 formulate하고 다양한 modality의 다른 hand input으로부터 학습하기 위한 세 가지 전략 제안

- latent hand space를 학습하기 위해 point-cloud 및 heat map과 같은 non-conventional 입력을 연구하고 RGB based hand pose estimation 시스템의 정확도를 향상시키기 위해 어떻게 활용할 수 있는지 보여준다. 이 논문에서 제안한 프레임 워크의 side-product는 RGB 이미지에서 실제처럼 보이는 포인트 클라우드를 합성 할 수 있다는 것이다.
- 공개적으로 사용 가능한 두 가지 벤치 마크로 평가하여 제안 된 프레임 워크가 훈련 중에 auxiliary modality를 최대한 활용하여 RGB pose estimation의 정확도를 높였다.
  chanllenging RHD 데이터셋에 대한 19%의 엄청난 향상을 포함하여 monocular RGB based hand pose estimation에서 우리의 estimated 포즈는 최신 방법을 능가한다.

Related Works

hand pose estimation approach를 분류하는 한 가지 방법

- generative
  - hand model를 employ 하고 hand model을 observation에 fit 하기 위해 optimization을 사용한다.
  - good initialization이 필요하며 그렇지 않으면 local minima에 빠질 수 있다.
- discriminative
  - visual observation을 hand pose에 직접 맵핑 하는 것을 학습 한다.
  - 대규모 annotated dataset로 인해, discriminative method를 기반으로 한 deep learning 이 매우 큰 성능을 낸다.

최근에 특히 depth나 3D 데이터를 입력으로 사용할 경우 더 정확하다. 그러나 3D 데이터는 training 혹은 testing시에 항상 available하지 않다. 몇몇 최근 작업은 monocular RGB data를 이용하기 시작했다. 가장 최근의 monocular RGB based 방법은 RGB 이미지로만 테스트를 수행하더라도 훈련을 위해 depth information을 활용한다.

Methodology

cross-modal method의 목적

- 어떤 다른 modality의 observation이 주어 졌을 때 target modality의 정보를 찾아 낼 수 있도록, 서로 다른 modality간의 관계를 찾아 내는 것.

cross modal VAE를 먼저 제시를 하고 multiple modality로 부터의 input과 output을 처리하기 위해 확장한다. 그리고 나서 두가지 latent space alignment operator strategy를 소개하고 이것을 RGB-based hand pose estimation에 적용하는 방법을 안내한다.

Cross Modal VAE and its extension

어떤 input modality로 부터 data sample x가 주어졌을 때, cross modal VAE는 latent variable z를 통해 evidence lower bound (ELBO)를 최대화 함으로써 target modality에서 coressponding target value y를 추정하는 것을 목표로 한다.

$$
\begin{aligned}
log\, p(y) \geq \mathrm{ELBO}_{\mathrm{cVAE}}(x; y; \theta, \phi) \qquad (1) \\
= E_{z \sim q_\phi} log\, p_\theta(y|z) - \beta D_{\mathrm{KL}}(q_\phi(z|x)||p(z))
\end{aligned}
$$

여기서 $D_{\mathrm{KL}}(\cdot)$ 는 Kullback-Leibler divergence, β 는 latent space capacity와 reconstruction accuracy의 balence를 맞추기 위한 hyper parameter 다.

$p(z) = \mathcal{N}(0, \mathbf{I})$ 은 latent variable z 의 Gaussian prior 다. variational approximation $q_\theta(z|x)$ 는 x 에서 z 로의 encoder 이고 $p_\theta(y|z)$ 는 z에서 y로의 디코더 혹은 inference network 다.

x 와 y 에 추가해서 N 개의 다른 modality $\{w_1, \ldots, w_N\}$ 로 부터 corresponding data가 있다고 가정하고 이런 modality는 latent representation z 가 주어졌을 때 조건부 독립이라고 가정한다.

단순하게 설명하기 위해 $N = 1$ 이라고 하자.

이 추가적인 modality를 encode 하기 위해서 식1에서 ELBO를 확장한다.

$$
\begin{aligned}
& log\, p(y, w_1) \geq \mathrm{ELBO}_{\mathrm{cVAE}}(x, w_1; y, w_1; \phi_{x,w_1}, \theta_y, \theta_{w_1}) \\
& = E_{z\sim\phi_{x,w_1}} log\, p_{\theta_y}(y|z) + \lambda_{w_1} E_{z\sim\phi_{x,w_1}} log\, p_{\theta_{w_1}}(w_1|z) - \beta D_{\mathrm{KL}}(q_{\phi_{x,w_1}}(z|x, w_1)||p(z)) \qquad (2)
\end{aligned}
$$

여기서 λ_{w_1} 는 $w1$ 과 y 사이에서 reconstruction accuracy를 regulate하는 hyperparameter 다.

original cross modal 과 더 많은 modality로 확장한 것에 대한 graphical model은 다음 <그림 2>에서 (a)와 (b)이다.

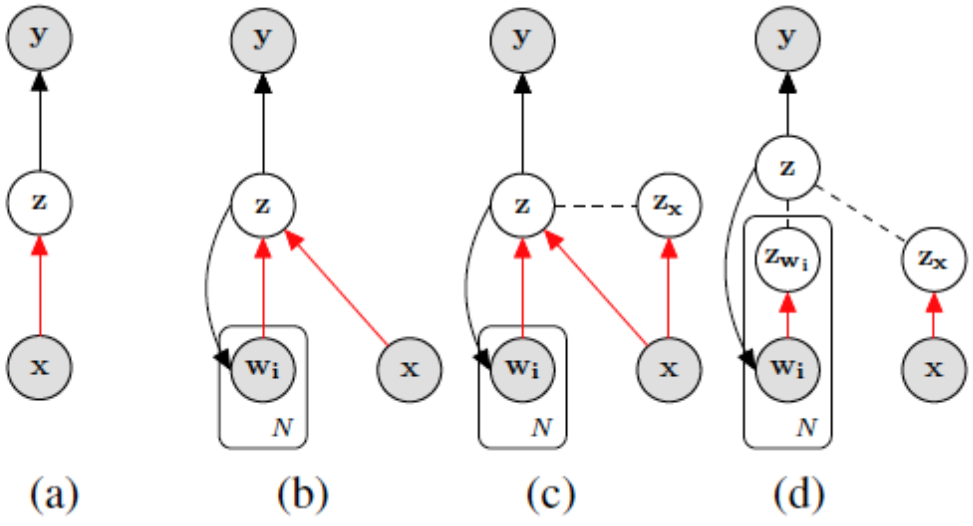


Figure 2: Graphical models. (a) Cross modal; (b) Extended cross modal; (c) Latent alignment with a KL divergence loss; (d) Latent alignment with the product of Gaussian experts. The shaded nodes represent observed variables while un-shaded nodes are latent. The red and black solid lines denote variational approximations q_ϕ or encoders, and the generative models p_θ or decoders respectively. The dashed lines denote the operation that embedding cross-modal latent spaces into a joint shared latent space; it is a KL divergence optimization for (c) and product of Gaussian experts for (d). Figure best viewed in colour.

<식 2>의 variational approximation $q_\phi(z|x, w_1)$ 로 부터 sampling 된 z는 <식 1>에서 $q_\phi(z|x)$ 로 부터 sampling된 것보다 더 informative 하기를 기대한다. 게다가, decoder $p_{\theta_{w_1}}$ 에 대한 expectation term 은 latent space 가 $y's$ modality에 over-fitting 되는 것을 막아 주기를 기대한다.

(여기서 부터 z를 z_{joint} 로 define한다.)

Algorithm 1 Extended cross modal with one encoder.

Require: $\mathbf{x}, \mathbf{y}, \mathbf{w}_1, T$
Ensure: $\phi_{\mathbf{x}}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
1: Initialize $\phi_{\mathbf{x}}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
2: **for** $t = 1, \ldots, T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})$
4: Decode $\mathbf{z}_{\mathbf{x}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_{\mathbf{x}}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_{\mathbf{x}})$
5: Update $\phi_{\mathbf{x}}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ via gradient ascent of $\text{ELBO}_{\text{cVAE}}(\mathbf{x}; \mathbf{y}, \mathbf{w}_1; \phi_{\mathbf{x}}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1})$
6: **end for**

위의 알고리즘 1을 보면 (식 2)는 실제로 응용하기 어려움이 있어 $\text{ELBO}_{\text{cVAE}}(x; y, w_1; \phi_x, \theta_y, \theta_{w_1})$

처럼 단순화 시켰다.

Latent Space Alignment

또 다른 방법은 $q_{\phi_x, w_1}(z|x, w_1)$ 와 $q_{\phi_x}(z|x)$ 를 결합해서 (jointly) 학습하고 두 분포를 함께 정렬하여 일치하는지 확인하는 것이다.

이 논문에서는 추론 능력을 향상 시키기 위해서 단일 modality로 부터 학습한 latent space를 joint modality로 학습한 latent space와 정렬하기 위해 joint training objective를 제안한다.

즉, modality w를 이용하기 위해 z_x를 z_{joint}와 align 하고 싶은 것이다. 이것은 $q_{\phi_x, w_1}(z|x, w_1)$ 와 $q_{\phi_x}(z|x)$ 를 최대한 가깝게 하려는 것이다.

KL divergence Loss

$$\begin{aligned}\mathcal{L}(\phi_{x,w_1}, \phi_x, \theta_y, \theta_{w_1}) &= \text{ELBO}_{\text{cVAE}}(x, w_1; y, w_1; \phi_{x,w_1}, \theta_y, \theta_{w_1}) \\ &+ \text{ELBO}_{\text{cVAE}}(x; y, w_1; \phi_x, \theta_y, \theta_{w_1}) \\ &- \beta\imath D_{\text{KL}}(q_{\phi_x, w_1}(z_{\text{zoint}}|x, w_1)||q_{\phi_x}(z_x|x)) \qquad (3)\end{aligned}$$

(식 3)은 인코딩 측면에서 2가지 문제점이 있다.

1. modality 혹은 N이 증가할 수록 joint encoder q_{ϕ_x, w_1} 은 학습하기 어려워 진다.
2. 단지, 2개의 인코더 q_{ϕ_x} 와 q_{ϕ_x, w_1} 를 가지고는 data pair (w_1, y)를 이용할 수 없다.

이런 문제점을 해결하기위 해 product of experts (PoE) 를 제안한다.

Product of Gaussian Experts

joint posterior는 각각의 posterior의 곱에 비례한다는 것이 입증되어 있다. 즉,

$$q(z|x, w_1) \propto p(z)q(z|x)q(z|w_1)$$

그 목적을 위해, 우리는 unimodal latent representation으로부터 joint latent representation을 추정할 수 있다.

VAE에서 $p(z)$와 $(q|z)$ 는 Gaussian 이다. Gaussian expert $q(z|x)$와 $q(z|w_1)$으로 부터 $q(|x, w_1)$을 도출할 수 있다. (그림 2d).

shared decoder의 도움을 받아, 다음 objective를 이용해서 joint latent representation에 도달한다.

$$\begin{aligned}\mathcal{L}(\phi_x, \phi_{w_1}, \theta_y, \theta_{w_1}) &= \mathrm{ELBO}_{\mathrm{cVAE}}(x; y; w_1; \phi_x, \theta_y, \theta_{w_1}) \\ &+ \mathrm{ELBO}_{\mathrm{cVAE}}(w_1; y, w_1; \phi_{w_1}, \theta_y, \theta_{w_1}) \\ &+ \mathrm{ELBO}_{\mathrm{cVAE}}(x, w_1; y, w_1; \phi_x, \phi_{w_1}, \theta_y, \theta_{w_1}) \\ &= E_{z_x \sim q_{\theta_x}}\ log\ p_\theta(y, w_1|z_x) + E_{z_{w_1} \sim q_{\phi_{w_1}}}\ log\ p_\theta(y, w_1|z_{w_1}) \\ &+ E_{z_{\mathrm{joint}} \sim \mathrm{GPod}(z_x, z_{w_1})}\ log\ p_\theta(y, w_1|z_{\mathrm{joint}}) \\ &- \beta(D_{\mathrm{KL}}(q_\theta(z_x|x)||p(z)) + (D_{\mathrm{KL}}(q_\theta(z_{w_1}|w_1)||p(z))) \qquad (4)\end{aligned}$$

여기서 $GProd(\cdot)$ 는 the product of Gaussian experts.

(식 3)의 KL divergence로 정렬 하는 경우 처럼 x와 w_1 의 joint encoder ϕ_{x,w_1} 는 필요하지 않다. 대신에 $q(z|x)$와 $q(z|w_1)$을 2개의 Gaussian Expert로 사용하였다.

$q(z|x) = \mathcal{N}(\mu_1, \Sigma_1)$, $q(z|w_1) = \mathcal{N}(\mu_2, \Sigma_2)$ 로 가정 하자. 그러면 Gaussian experts 는 product는 mean μ 과 covariance Σ 를 갖는 Gaussian 이다.

$$\begin{aligned}\mu = (\mu_1 T_1 + \mu_2 T_2)/(T_1 + T_2), \\ \sigma = 1/(T_1 + T_2) \\ \text{여기서} \quad T_1 = 1/\Sigma_1, T2 = 1/\Sigma_2\end{aligned}$$

Algorithm 2 Latent alignment with Eq. 3.

Require: $\mathbf{x}, \mathbf{y}, \mathbf{w}_1, T$

Ensure: $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$

1: Initialize $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
2: **for** $t = 1, \ldots, T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})$
4: Encode $\mathbf{x}, \mathbf{w}_1$ to $q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}_{\mathrm{joint}}|\mathbf{x}, \mathbf{w}_1)$
5: Decode $\mathbf{z}_{\mathbf{x}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_{\mathbf{x}}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_{\mathbf{x}})$
6: Decode $\mathbf{z}_{\mathrm{joint}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_{\mathrm{joint}}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_{\mathrm{joint}})$
7: Construct $D_{KL}(q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}_{\mathrm{joint}}|\mathbf{x}, \mathbf{w}_1)||q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x}))$
8: Update $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ via gradient ascent of Eq. 3
9: **end for**

Algorithm 3 Latent alignment with Eq. 4.

Require: $\mathbf{x}, \mathbf{y}, \mathbf{w}_1, T$

Ensure: $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$

1: Initialize $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
2: **for** $t = 1, \ldots, T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})$
4: Encode $\mathbf{w}_1$ to $q_{\phi_{\mathbf{w}_1}}(\mathbf{z}_{\mathbf{w}_1}|\mathbf{w}_1)$
5: Construct $\mathbf{z}_{\text{joint}} = \text{GProd}(\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{w}_1})$
6: Decode $\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{w}_1}, \mathbf{z}_{\text{joint}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\cdot), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\cdot)$ respectively
7: Update $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ via gradient ascent of Eq. 4
8: **end for**

Impementation Details

Data Pre-processing and Augmentation

- RGB 이미지에서 hand를 포함하는 영역을 ground truth mask를 이용해서 crop하고 256x256으로 resizing 한다.
- depth image에서 대응되는 영역은 camera intrinsic parameter를 사용해서 point cloud로 변환된다. 카메라의 내부 파라미터로는 초점거리(focal length), 주점(principal point), 비대칭계수(skew coefficient) 같은 것들이 있다.
- 각 training step마다 training input 으로서 256개의 다른 point가 샘플링 된다.

Viewpoint correction

- RGB 이미지로 부터 hand를 cropping 한 후에 image 내에서의 hand center를 이미지의 중앙으로 이동시킨다.

Data augmentation

- scaled randomly between [1,1.2]
- translated [-20,20] pixels
- camera view axis를 중심으로 rotated [$-\pi$,π]
- hue of image 는 [-0.1,01] 구간 내에서 랜덤으로 조정
- point cloud는 camera view axis를 중심으로 랜덤하게 회전
- 3D pose label도 그것에 따라서 회전

Encoder and Decoder Modules

- 이 작업에서는 RGB와 point cloud 에 대한 encoder를 학습하고 3D hand pose, point cloud 그리고 RGB image의 2D hand key point의 heat map에 대한 decoder를 학습한다.
- RGB image encoder: Resnet-18. latent variable의 mean과 variance vector를 예측하기 위해 2개의 추가적인 Fully connected layer를 사용
- point cloud encoder : ResPEL network (연산부하 때문에 hidden unit의 수를 절반으로 줄임)
- heat map decoder: DC-GAN의 decoder architechture
- heat map decoder loss function

$$\mathcal{L}_{\text{heat}} = \sum_{j-1}^{J} \|\hat{H}_j - H_j\| \qquad (7)$$

$$H_j \ : \ \text{ground-truth heatmap for the} j\text{-th hand keypoint}$$

$$\hat{H}_j \ : \ \text{the prediction}$$

- point cloud decoder: FoldingNet architecture
- point cloud decoder loss function
  - **Chamfer distance** : the sum of the Euclidean distance between points from one set and its closest point in the other set

$$\mathcal{L}_{\text{Chamfer}} = \frac{1}{|P|} \sum_{p \in P} min_{\hat{p} \in \hat{P}} \|\hat{p} - p\| + \frac{1}{|\hat{P}|} \sum_{\hat{p} \in \hat{P}} min_{p \in P} \|\hat{p} - p\| \qquad (8)$$

- **Earth Mover's disstance (EMD)** : one-to-one bijective (전단사) correspondences are established between two point clouds, and the Euclidean distances between them are summed.

$$\mathcal{L}_{\text{EMD}} = min_{\phi : P \to \hat{P}} \frac{1}{|P|} \sum_{p \in P} \|p - \phi(p)\| \qquad (9)$$

- 3d pose decoder: 4 fully-connected layers with 128 hidden units for each layer.
- 3d pose decoder loss function

$$\mathcal{L}_{\text{pose}} = \|\hat{y} - y\| \qquad (10)$$

- **reconstruction loss functions**

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{pose}} + \lambda_{\text{heat}} \mathcal{L}_{\text{heat}} + \lambda_{\text{cloud}} (\mathcal{L}_{\text{Chamfer}} + \mathcal{L}_{\text{EMD}}) \qquad (11)$$

- $$\text{overall loss} = \mathcal{L}_{\text{recon}} + D_{\text{KL}}$$

Experimentation

- the dimensionality of latent variable z : 64
- λ_{heat} : 0.01
- λ_{cloud} : 1
- β' : 1
- Tensorflow

- Adam optimizer (initial learning rate : 10^{-4})
- learning rate는 수렴 후 10을 지수로 2배씩 낮춘다.
- batch size : 32
- β : from 10^{-5} to 10^{-3}

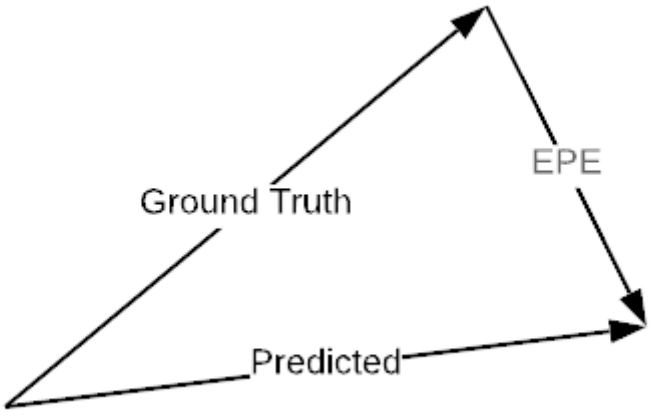
dataset and evaluation metrics

데이터 셋

- Rendered Hand Pose Dataset (RHD)
    - RHD는 39개의 동작을 수행하는 20명의 characters로부터 320×320 해상도인 렌더링된 손 이미지의 합성 데이터 세트
    - training : 41238개
    - testing : 2728개
    - 각 RGB 이미지에는 대응되는 depth map, segmentation mask 그리고 3D hand pose가 제공된다.
    - 이 데이터 세트는 다양한 시각적 풍경, 조명, 소음 때문에 매우 어렵다.
- Stereo Hand Pose Tracking Bechmark (STB)
    - 한 사람의 왼손이 여섯 개의 서로 다른 실제 배경 앞에 있는 비디오를 포함하고 있다.
    - 데이터 세트는 스테레오 이미지, 640 × 480 해상도의 color-depth pairs, 3D hand pose annotations 을 제공한다.
    - 데이터 세트의 12개 시퀀스 각각에 1500개의 프레임이 포함되어 있다. (10개는 train용 , 2개는 test용으로 사용함.)

metrics

- mean end-point-error (EPE)



- area under the curve (AUC) on the percentage of correct keypoints (PCK)

Qualitative results

some qualitative examples of poses and point clouds decoded from the z_{rgb}

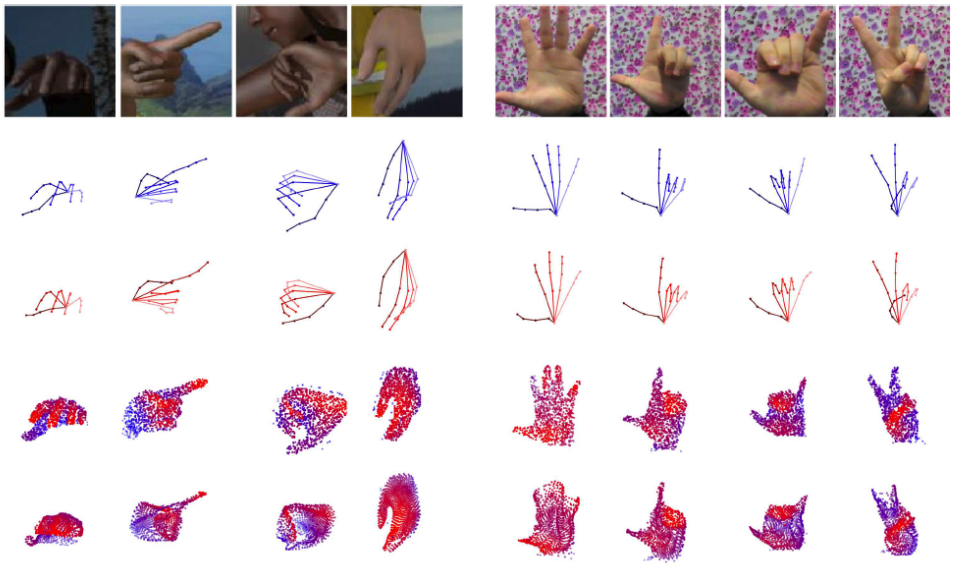


Figure 3: 3D pose estimation and point cloud reconstruction for RHD (left) and STB (right) dataset. From top to bottom: RGB images, ground-truth poses in blue, estimated poses from $\mathbf{z}_{rgb}$ in red, ground-truth point clouds, reconstructed point clouds from $\mathbf{z}_{rgb}$. The color for point clouds decodes the depth information, closer points are more red and further points are more blue. Note that the ground-truth point clouds are not used for inference, it is shown here only for comparison purpose.

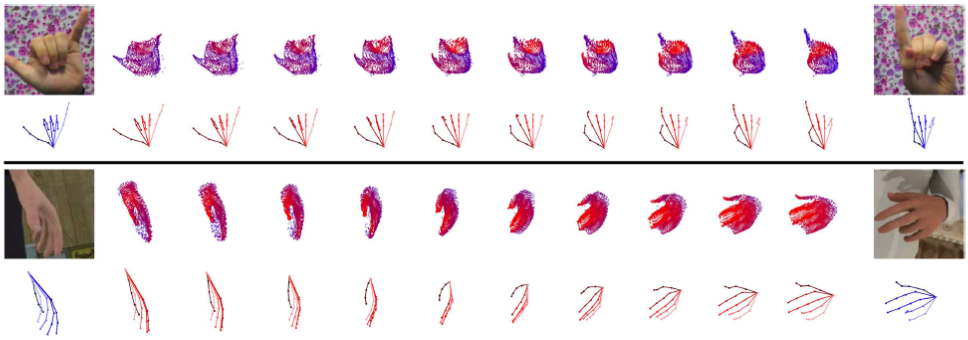


Figure 4: Latent space interpolation. Two examples of reconstructing point clouds and hand poses from the latent space. The most left and most right column are RGB images and their corresponding ground-truth poses. Other columns are generated point clouds and poses when interpolating linearly on the latent space.

RGB 3D Hand Pose Estimation

Strategy	Encoder	Decoder	Mean EPE [mm]
S1 (Eq. 1)	R	P	16.61
S2 (Alg. 1)	R	H+P	16.10
	R	C+P	15.91
	R	C+H+P	15.49
S3 (Alg. 2)	R+C	C+H+P	14.93
S4 (Alg. 3)	R+C	C+H+P	**13.14**

Table 1: Comparison of different training strategies on the RHD dataset. The mean EPE values are obtained from monocular RGB images. (R: RGB, C: point cloud, P: pose, H: heatmap). Poses estimated from monocular RGB images can be improved by increasing number of different encoders and decoders during training.

Viewpoint correction

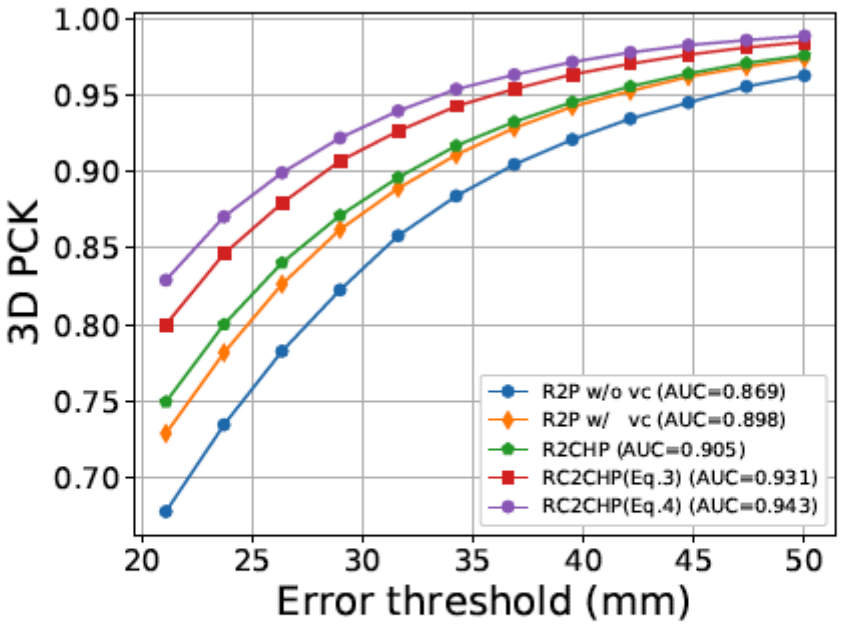


Figure 5: Comparisons of 3D PCK results of our different strategies on RHD dataset. The abbreviations can be found in Sec. 3.3 and "vc" stands for "view correction"

Comparison to state-of-the-art

	Method	RHD	STB
VAE-based	Spurr *et al.* [19]	19.73	8.56
	Yang *et al.* [27]	19.95	8.66
	Ours	**13.14**	**7.05**
Others	Z&B [31]	30.42	8.68
	Iqbal *et al.* [9]	13.41	\

Table 2: Comparison to state-of-the-art on the RHD and STB with mean EPE [mm]. Ours refers to S4 in Table 1 (RC2CHP).

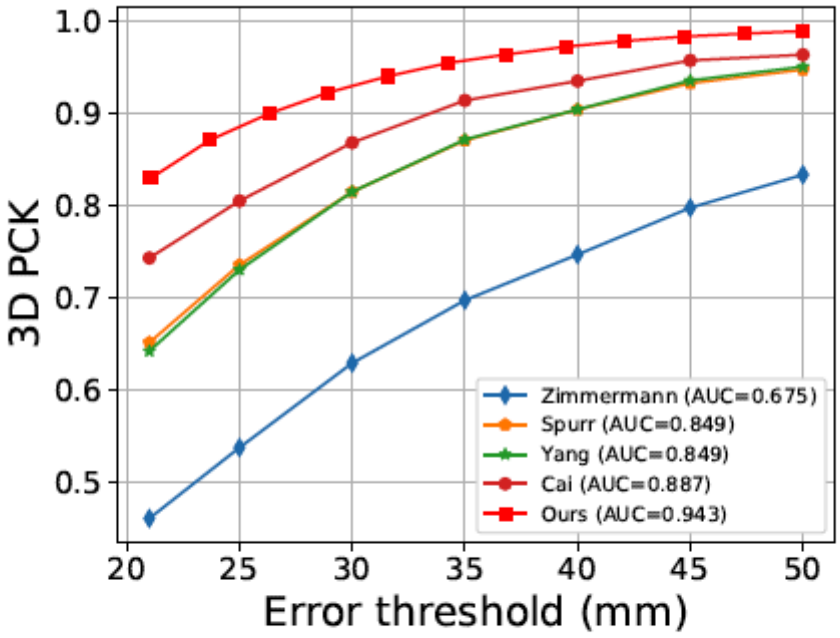


Figure 6: AUC: Comparison to state-of-the-art methods on the RHD dataset. Ours refers to S4 in Table 1 (RC2CHP).

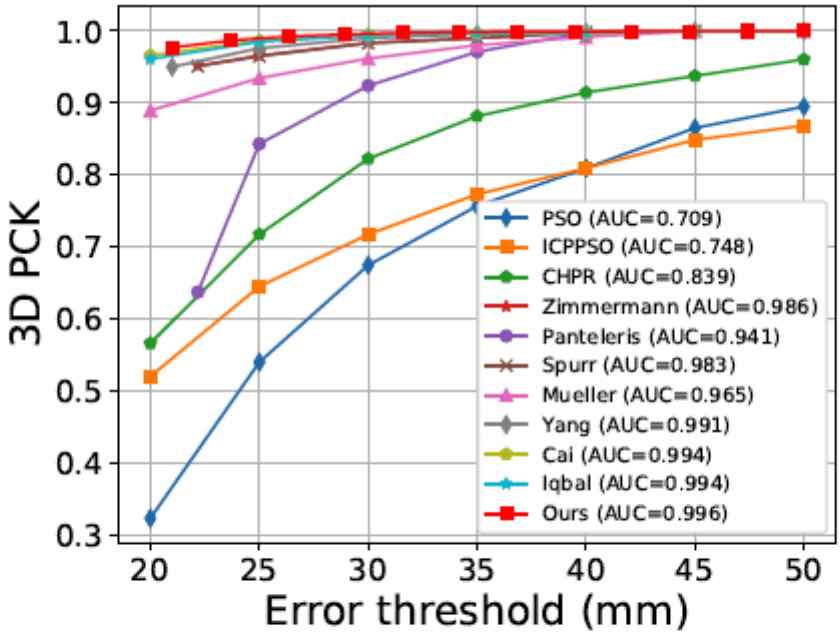


Figure 7: AUC: Comparison to state-of-the-art methods on the STB dataset. Ours refers to S4 in Table 1 (RC2CHP).

Weakly-supervised learning

- 제안된 방법의 flexibility로 인해 point cloud는 training을 도와주기 위해 unlabelled data에 대한 "weak" 라벨로 사용할 수 있다.
- 처음 m% 샘플은 라벨된 데이터로 사용하고 (RGB, point cloud 그리고 3D pos 포함) , 나머지는 unlabelled data로 사용(라벨링 때문에 3D pose는 사용하지 않음)해서 테스트를 했다.

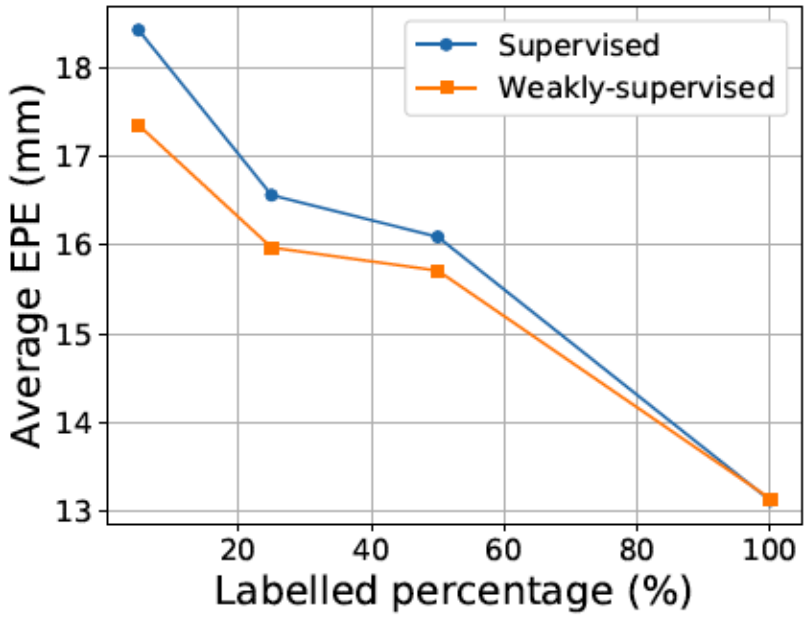


Figure 8: Mean EPE of our model on the weakly-supervised setting. Our method makes full use of unlabelled data, as the weakly-supervised setting performs almost as well as the supervised one.

결론

- 이 논문에서는 RGB-based hand pose estimation을 multimodal 학습과 cross-modal inference로 formulate했다.
- 3가지 hand modality에 대해 각기 다른 목표를 도출하고 관련된 latent space를 하나로 합쳐서 하나로 정렬 하는 각기 다른 방법을 보여줬다.