3D Hand Pose Estimation via aligned latent space injection and kinematic losses

Andreas Stergioulas* Theocharis Chatzis* Dimitrios Konstantinidis
Kosmas Dimitropoulos Petros Daras, $SeniorMember, IEEE$
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
{andrster, hatzis, dikonsta, dimitrop, daras}@iti.gr

Abstract

In this paper, we propose a novel multi-stage deep learning methodology to accurately tackle the problem of hand pose estimation. More specifically, we initially propose a disentanglement stage to differentiate the significant pose-specific information from the irrelevant background noise and illumination variations of RGB images. Then, we introduce a variational alignment stage to accurately align the latent spaces of the pose-specific and the true hand pose information, effectively improving the discrimination ability of the proposed methodology. Finally, we propose the use of two loss terms to impose physiological and geometrical kinematic constraints to the predicted hand poses, empowering the proposed methodology to avoid non-plausible poses. During all stages, a novel injection decoder is introduced, improving significantly the decoding accuracy of the latent space. Extensive experimentation on two well-known datasets (i.e., RHD and STB) validate the ability of the proposed methodology to achieve accurate hand pose estimation results, overcoming current state-of-the-art methods.

1. Introduction

3D hand pose estimation involves the prediction of the position and orientation of the hand and fingers relative to a coordinate system, given an RGB or depth image. It plays a crucial role in a wide range of application fields, such as gesture recognition [1, 17], augmented reality (AR) [18, 29], virtual reality (VR) [3, 5], avatar animation [16, 35] and human computer interaction (HCI) [21, 34]. Moreover, 3D hand pose estimation can be beneficiary to the Deaf community, by being incorporated in automated sign language recognition and translation systems [2, 15, 27, 35].

Conventionally, 3D hand pose estimation is performed by processing depth images [25, 26, 37, 45]. The employment of RGB images as input data for 3D hand pose estimation has recently started to gain attention [12, 32, 33, 36, 41, 47], due to the availability of large annotated RGB datasets and the recent advances in Deep Neural Networks (DNNs). However, leveraging RGB data for 3D hand pose estimation can be really challenging due to the inherent ambiguities that are usually present in RGB images, such as difficulties in depth estimation, self-occlusions and background and illumination variations. Such ambiguities can significantly downgrade the accuracy of 3D hand pose estimation methods, therefore, the use of deep neural networks with high generalization abilities is imperative.

To this end, recent hand pose estimation methods utilize generative networks, such as Variational Autoencoders (VAEs) [14] and Generative Adversarial Networks (GANs) [8] that demonstrate tremendous learning capacity and generalization capabilities. VAEs and GANs are capable of constructing highly descriptive latent spaces that can accurately model input data and generalize on unseen data.In this context, literature works [33, 36] aim at aligning the RGB latent space with the ground truth pose latent space, improving the accuracy of hand pose estimation models. Nevertheless, the presence of the RGB context information (e.g., illumination and background variations) can severely degrade the performance of such methods. To this end, recent works [41, 9] attempt to differentiate the meaningful hand pose information from the irrelevant RGB context information by proposing new latent subspaces. However, the accurate extraction of the meaningful pose information is a challenging task, due to the the highly entangled nature of the RGB images, while there is also the issue of restricting the predicted poses to the space of humanly plausible ones.

Motivated by the need for an accurate alignment between the RGB image space and the ground truth 3D hand pose information, we propose a novel multi-stage methodology to accurately infer the 3D hand poses from RGB images. At the first stage, we employ an adversarial network to disentangle the input RGB images into the pose-specific and the RGB context latent subspaces, thus discriminating the significant from the irrelevant information. At the second stage, we employ variational mappers to align the pose-

*equal contribution

specific and the ground truth pose latent spaces, enhancing in this way the prediction performance of the pose-specific latent space. At the third stage, we use two loss terms that impose physiological and geometrical constraints on the predicted hand poses to avoid non-plausible ones. Finally, this work proposes a novel injection decoder that is employed during all stages to boost the discrimination ability (especially during disentanglement) and the decoding accuracy of the constructed latent spaces. To summarise, the main contributions of this work are:

- We introduce a novel multi-stage hand pose estimation methodology that achieves improved cross-modal alignment between RGB images and ground truth 3D hand poses. A GAN network initially disentangles the meaningful pose-specific information from the RGB images before variational mappers align the extracted pose-specific latent subspace with the true 3D hand pose latent space.

- We use two loss terms to refine the predicted hand poses by imposing physiological and geometrical kinematic constraints. The first loss is based on the well-known KCS representation [38, 39] and aims to regulate the predicted poses based on the length of the hand bones, whereas the second loss imposes restrictions on the relative position among the joints.

- We implement a novel injection decoder that can be employed in any VAE framework to enhance the disentanglement process, the discrimination ability of the model and the decoding accuracy of the latent space. The proposed decoder uses residual connections to pass the latent space sample to the intermediate layers of the decoder, improving the gradient flow.

- We conduct thorough experiments on two well-known publicly available RGB datasets, RHD [47] and STB [42], showcasing the superiority of the proposed methodology against other state-of-the-art RGB-based hand pose estimation methods.

2. Related Work

Methods on 3D hand pose estimation can be classified into the following categories, depending on the employed input modalities [6]: (i) Depth-based [20, 22, 30], (ii) RGB-based [33, 36, 44] and (iii) Multimodal [7, 40, 46] ones. In addition, hand pose estimation approaches can be further subdivided into (a) Model-based [4, 11, 44] and (b) Model-free [12, 33, 36] methods. Model-based approaches attempt to fit the detected 3D hand points to a predefined hand model (i.e., define the hand model parameters), while Model-free approaches output raw 3d hand points. In this work, we only focus on Model-free RGB-based approaches, thus the remaining section reviews related previous works that fall under this category.

Zimmerman et al. [47] introduced one of the first deep-learning based methods to estimate 3D hand joint locations from RGB images. Their framework consisted of three elementary units: the first unit was responsible to locate the hand region (HandSegNet), the second unit generated score maps for each 2D keypoint (PoseNet) and the last unit regressed 3D joint locations from the predicted 2D ones.

On the other hand, Iqbal et al. [12] presented a novel 2.5D pose representation. They employed an Hourglass network [24] to extract latent 2D heatmaps and depth maps for each keypoint in order to better address the depth ambiguities residing in RGB images. In a similar manner, Moon et al. [23] employed a ResNet network [10] to extract image features and two upsamplers to estimate a 2.5D pose representation for each hand, which then used to reconstruct the 3D hand pose.

Identifying the need for better generalization in hand pose estimation, several authors employed generative networks for their excellent generalization and descriptive capabilities. Spurr et al. [33] proposed a cross-modal VAE-based framework to regress 3D hand joint locations from RGB images by iteratively training encoder-decoder independent pairs. Theodoridis et al. [36] introduced a novel multi-stage variational framework to map the latent spaces generated from RGB-to-Pose and Pose-to-Pose VAE networks. Initially, they trained the above networks independently prior to the use of a VAE mapper component to map the cross-modal latent space to the more descriptive single-modal one.

While the previous works modeled the entire RGB image into a latent space distribution, other works engaged in identifying the meaningful 3D hand pose information and differentiating it from the RGB context. Yang et al. [41] presented a disentangled VAE network (dVAE) that created a common latent space. They learned disentangled representations of RGB images and 3D keypoint locations, enabling for specific sampling and inference of various factors, such as background, viewpoint, etc. Gu et al. [9] proposed a VAE-based framework to disentangle the 3D hand pose from the latent space. To achieve this, the authors applied adversarial training to bisect each of the RGB-to-RGB and Pose-to-Pose spaces into two subspaces and a network consisting of fully connected layers to translate the modality-context subspaces. To encourage the predicted 3D hand poses to be more anatomically valid, Spurr et al. [32] proposed a weakly supervised 3D hand pose estimation approach that correct the joint predictions using biomechanical constraints.

In this paper, we propose a novel cross-modal alignment method to create a pose-specific latent space that is well-separated from the RGB context and fully aligned with the

latent space of the true hand poses. The accuracy of the proposed method is further boosted by two kinematic losses that refine the predicted hand poses and a novel injection decoder that improves the decoding of the latent spaces.

3. Methodology

The proposed hand pose estimation methodology aims at achieving an improved alignment between the RGB images and the ground truth hand poses. A GAN architecture is initially employed to disentangle the relevant pose-specific information of the RGB images from the irrelevant RGB context, effectively limiting its impact on the predictions. Then, two variational mappers are responsible for bringing the disentangled pose information closer to the ground truth 3D hand pose information. Finally, kinematic physiological and geometrical constraints are introduced to ensure that the predicted hand poses are plausible and there are no finger deformations. A novel injection decoder is also introduced to the proposed VAE network architectures to inject the latent space sample at each decoder layer through residual connections, thus creating a more discriminative latent space (especially during the disentanglement stage) and boosting the decoding accuracy of the VAE networks. An illustration of the proposed multi-stage hand pose estimation methodology is presented in Figure 1.

3.1. Cross-modal alignment

This section describes the novel cross-modal alignment strategy that consists of the disentanglement stage, aiming at differentiating the RGB context information from the pose-specific information, and the variational alignment stage, aiming at aligning the pose-specific and the true 3D hand pose latent spaces.

3.1.1 Disentanglement stage

The first stage aims at disentangling the information embedded in the input data (i.e., RGB images) to two latent subspaces that model the significant pose-specific information and the irrelevant RGB context information, respectively. In this way, we aim at improving the robustness of the proposed hand pose estimation method to the effects of illumination variations and background colour and texture.

To this end, an RGB-to-Pose VAE is trained alongside a Pose-to-Pose VAE, while a discriminator is employed to disentangle, in an adversarial manner, the latent space of the RGB-to-Pose VAE. More specifically, the cross-modal encoder E_{RGB}, encodes input images $\boldsymbol{X}$ into pairs of mean and variance fixed-vectors, $(\mu_{cont}^{RGB}, \sigma_{cont}^{RGB})$ and $(\mu_{pose}^{RGB}, \sigma_{pose}^{RGB})$ that model the RGB context and pose-specific information, respectively. These vectors approximate two Gaussian distributions for each disentangled latent subspace. On the other hand, the uni-modal encoder, E_{pose}, encodes the 3D joint coordinates $\boldsymbol{J} \in \mathbb{R}^{N\times 3}$ into $(\mu_{pose}^{true}, \sigma_{pose}^{true})$, generating the true posterior distribution of the 3D hand poses z_{pose}^{true}.

Subsequently, a generative adversarial network is formed, in which the cross-modal encoder E_{RGB} acts as the generator that produces plausible hand poses, while the discriminator Dis tries to distinguish between the predicted poses and the ground truth poses. In this adversarial context, the discriminator is trained to bridge the gap between the pose-specific z_{pose}^{RGB} and the unimodal z_{pose}^{true} latent spaces, thus effectively isolating the factors related to the hand pose from the latent space of the RGB image information. The discriminator's objective is therefore:

$$L_{Dis} = 1/2(L_{BCE}(Dis(z_{pose}^{RGB}), 0)) + 1/2(L_{BCE}(Dis(z_{pose}^{true}), 1)), \quad (1)$$

where L_{BCE} is the common binary cross entropy loss.

Afterwards, a sample is stochastically drawn from the pose-specific latent subspace z_{pose}^{RGB} and fed into ID_{RGB} to infer the 3D joint coordinates. As far as the uni-modal VAE is concerned, a sample is passed from the z_{pose}^{true} to ID_{pose}.

In order to train end-to-end the disentangled framework, the common VAE loss is optimized for both the cross-modal and the uni-modal VAE, where:

$$L_{VAE}^{cross} = L_{MSE}(J, ID_{RGB}(z_{pose}^{RGB})) - \beta_{VAE}^{cross} L_{KL}(E_{RGB}(z_{cont}^{RGB}|\boldsymbol{X})||p(z)) - \beta_{VAE}^{cross} L_{KL}(E_{RGB}(z_{pose}^{RGB}|\boldsymbol{X})||p(z)) \quad (2)$$

is the cross-modal VAE objective and

$$L_{VAE}^{uni} = L_{MSE}(J, ID_{pose}(z_{pose}^{true})) - \beta_{VAE}^{uni} L_{KL}(E_{pose}(z_{pose}^{true}|\boldsymbol{J})||p(z)) \quad (3)$$

is the uni-modal VAE objective. The terms L_{MSE}, L_{KL} denote the *MSE* loss and the *Kullback–Leibler divergence* [14], respectively. The parameters β control the weight of the KL divergence. The overall objective of our model during this training stage is formulated as:

$$L_{disentangle} = L_{VAE}^{cross} + L_{VAE}^{uni} + L_{Dis} \quad (4)$$

3.1.2 Variational alignment stage

This stage is concerned with finding an effective way to align the disentangled pose-specific and the uni-modal latent spaces, since the latest contains more accurate information about the 3D hand poses. In this way, we aim at leveraging the ground truth pose information to construct a highly descriptive RGB latent space to achieve improved hand pose estimation results.

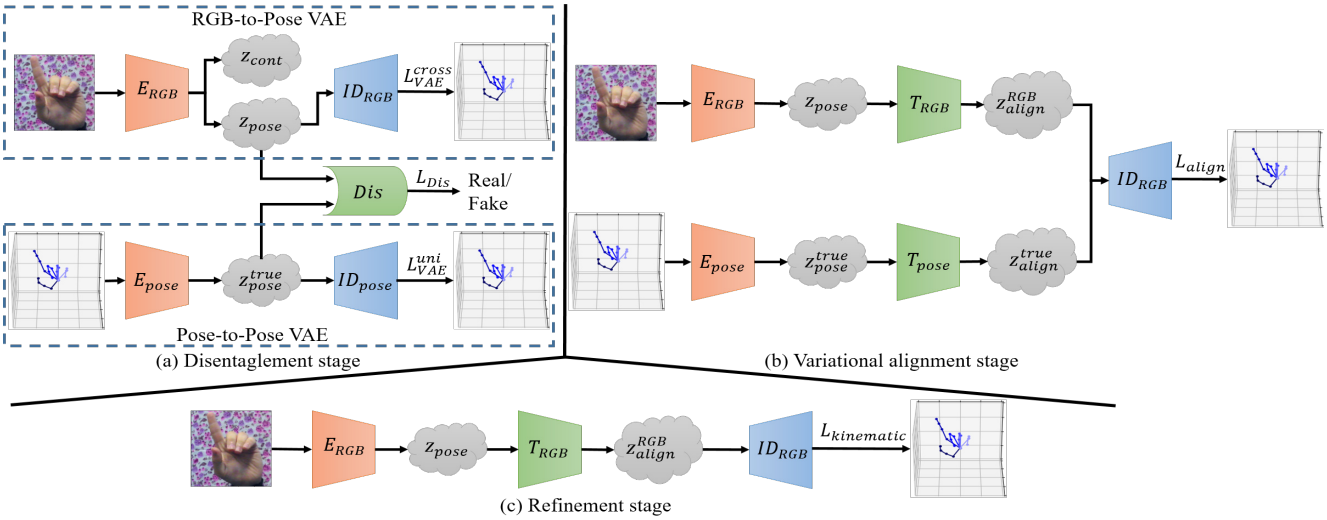


Figure 1: An overview of the proposed multi-stage hand pose estimation methodology. (a) The disentanglement stage extracts the meaningful pose-specific information from the RGB images. (b) The variational alignment stage aligns the pose-specific and ground truth pose latent spaces. (c) The refinement stage imposes kinematic constraints to the predicted poses.

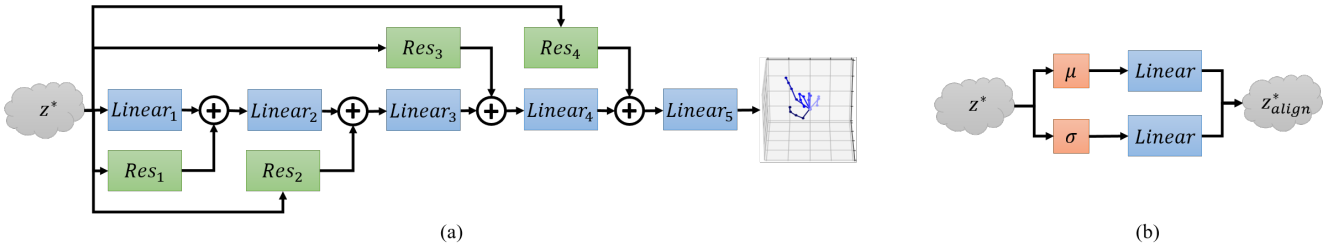


Figure 2: Network architectures of a) the injection decoder and b) the variational mapper. With z^* we denote any initial latent space, while with z^*_{align} we denote any aligned latent space. Res denotes a linear layer used as a residual connection.

To this end, we propose the use of two variational alignment components, T_{RGB} and T_{pose}, to project each latent space to a new one. By jointly training them while using a shared decoder, an efficient alignment of the two latent spaces is achieved. The architecture of both alignment components is the same and can be viewed in Figure 2b. T_{RGB} takes the vector pair produced by E_{RGB}, $(\mu_{pose}^{RGB}, \sigma_{pose}^{RGB})$, as input and applies a re-parameterization, generating a new distribution with $(\mu_{align}^{RGB}, \sigma_{align}^{RGB})$. Likewise, T_{pose} is responsible to generate a new distribution with $(\mu_{align}^{true}, \sigma_{align}^{true})$, given $(\mu_{pose}^{true}, \sigma_{pose}^{true})$. Afterwards, the pretrained ID_{RGB} decoder is utilized to infer the 3D hand pose from each latent space.

Besides the joint regression task, we need to regularize the aligned latent spaces in order to facilitate the generative process, since the alignment is based on variational inference. This is accomplished by using the Kullback–Leibler divergence alongside the MSE loss. Thus, we optimize the network's parameters using the following objective:

$$L_{align} = L_{align}^{cross} + wL_{align}^{uni}, \tag{5}$$

where:

$$L_{align}^{cross} = L_{MSE}(J, ID_{RGB}(z_{align}^{RGB})) - \beta_{align}^{cross} L_{KL}(T_{RGB}(z_{align}^{RGB}|\boldsymbol{X})||p(z)) \tag{6}$$

and

$$L_{align}^{uni} = L_{MSE}(J, ID_{RGB}(z_{align}^{true})) - \beta_{align}^{uni} L_{KL}(T_{pose}(z_{align}^{true}|\boldsymbol{J})||p(z)) \tag{7}$$

The term w is a weight that controls the contribution of the uni-modal VAE loss term. The VAE encoders E_{RGB} and E_{pose} are initially frozen while the two alignment components and the shared decoder are trained. After convergence, a finetuning process is performed where the entire

network is trained, leading to further alignment enhancement.

3.2. Refinement stage with kinematic losses

During the refinement stage, two loss terms are employed, namely the Kinematic Chain Space (KCS) and the Geometrical (GEO) losses. The aim of these loss terms is to impose kinematic constraints on the predicted 3D hand poses based on physiological and geometrical criteria in order to discard non-plausible poses and finger deformations. The physiological constraints are derived from projecting the 3D hand joints to a kinematic chain [38], aiming to minimize the discrepancy between the ground truth and the predicted hand poses, whereas the geometrical constraints are based on the calculation of the joint-line distances of hand joints [43] and minimizing the error from the ground truth distances.

3.2.1 KCS Loss

The kinematic chain space has been employed in the literature [38, 39] as an alternative way of 3D skeleton representation that contains joint angles and bone lengths. Leveraging this representation, we propose a loss term that is directly applied to the kinematic chain space. More specifically, a bone b_k is the vector between the r-th and t-th joint,

$$b_k = p_r - p_t = Jc, \tag{8}$$

with c being a vector with a value of 1 at position r and -1 at position t and $J \in \mathbb{R}^{N\times 3}$ the 3D joint coordinates

$$c = (0, ..., 0, 1, 0, ..., 0, -1, 0, ..., 0)^T \tag{9}$$

A matrix $B \in \mathbb{R}^{3\times b}$ containing all the hand bones, can be constructed by concatenating b bones:

$$B = (b_1, b_2, ..., b_b) \tag{10}$$

Moreover, the matrix $C \in \mathbb{R}^{j\times b}$ is constructed by concatenating c vectors. Therefore, the B matrix is calculated as follows:

$$B = JC \tag{11}$$

The KCS is computed by multiplying B with its transpose:

$$KCS \quad = \quad B^T B \quad = \quad \begin{bmatrix} l_1^2 & & & \\ & l_2^2 & & \\ & & \ddots & \\ & & & l_b^2 \end{bmatrix} \tag{12}$$

Each entry in KCS contains the inner product of two bone vectors and a scaled angular representation on the other entries [39]. The KCS loss is then formulated as:

$$L_{KCS} = \ell_1(KCS^{groundtruth}, KCS^{predicted}), \tag{13}$$

with ℓ_1 being the L1 distance.

3.2.2 Geometrical Loss

The proposed geometrical loss term is based on the calculation of joint-line distances, which can be considered as an additional spatial representation that models the relationship among the 3D joints of a hand. Given three different joints i, j, k and the distances between them d_{ij}, d_{ik}, d_{jk}, the joint-line distance of joint i to the line l formed by joints j and k is equal to the shortest (i.e., perpendicular) distance between the point and the line and it can be efficiently computed using the Heron's formula as shown below:

$$S_\Delta(i, l) = 2\frac{\sqrt{s(s - d_{ij})(s - d_{ik})(s - d_{jk})}}{d_{jk}}, \tag{14}$$

where $s = 0.5(d_{ij} + d_{ik} + d_{jk})$. Following the above procedure, 190 lines are computed for the 21 hand joints. Then, a matrix of joint-line distances is formed that is equal to:

$$Dist_{J,L} = \sum_{i}^{J}\sum_{l}^{L} S_\Delta(i, l) \tag{15}$$

Using the L1 distance between the ground truth joint-line distances and the joint-line distances of the predicted joints, the geometrical loss is estimated as:

$$L_{GEO} = \ell_1(Dist_{J,L}^{groundtruth}, Dist_{J,L}^{predicted}) \tag{16}$$

The final loss during the kinematic training stage is formulated as:

$$L_{kinematic} = L_{MSE}(J, J^{predicted}) + qr_{kcs}L_{KCS} + q'r_{geo}L_{GEO}, \tag{17}$$

with q taking values between 0 and 1, alternating between L_{KCS} and L_{GEO} and r_{kcs}, r_{geo} being hyperparameters that control the weights of the loss terms.

3.3. Injection Decoder

Differentiating from the literature that considers decoders as stacks of fully connected layers and motivated by residual connections, this work proposes the novel injection decoder that consists of fully connected layers with intermediate residual layers. The proposed decoder functions inside any VAE framework and is present in our network architecture during the entire multi-stage training process. Its purpose is to improve the discrimination ability and the performance of a VAE network by allowing the construction of more descriptive latent spaces and by enabling a better flow of gradients during training.

More specifically, to learn a variational mapping from modality x to modality y, an encoder E_i is initially employed to project the input $x \in \{X\}$ into the latent space LS_i by producing fixed-size vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ with dimensionality d_i, which approximates a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = diag(\sigma_i(1)^2, \ldots, \sigma_i(d_i)^2)$. A

decoder D_i then, draws stochastically a sample $\boldsymbol{z}$ from the distribution and decodes it to $y \in \{Y\}$. In the proposed injection decoder ID, the sample $\boldsymbol{z}$ is not only provided at input level but also injected at every intermediate layer by performing an element-wise addition between the output of the layer and the initial $\boldsymbol{z}$, as shown in Figure 2a. Formally, for K intermediate layers, this can be expressed as:

$$y_i^{inter} = \zeta(\sigma(W_i y_{i-1}^{inter})) + \zeta(\sigma(W_i^s z)), i \in [2, ..., K-1], \quad (18)$$

where W_i denotes the layer's learned weight matrix, $W_i^s z$ denotes a learned matrix that projects $\boldsymbol{z}$ to the same dimension as y_{i-1}^{inter}, σ represents batch normalization and ζ represents a ReLU activation function. The biases are omitted for simplification purposes. If $i = 1$, the left term of the addition in Equation 18 becomes $\zeta(\sigma(z))$ and if $i = K$, the injection decoder predicts the 3D position of N hand joints, without the need for a residual connection:

$$J^{predicted} = \zeta(\sigma(W_i y_{K-1}^{inter})), y^{out} \in \mathbb{R}^{N \times 3} \quad (19)$$

4. Experiments

4.1. Datasets and Metrics

The proposed method is tested on two publicly available benchmark datasets, namely Stereo Hand Pose Tracking (STB) [42] and Rendered Hand Pose (RHD) [47] datasets.

STB is a real dataset that comprises one subject performing 12 hand motion sequences with 6 different backgrounds. Altogether it contains 18K frames with 640×640 resolution and a 15K/3K training/test split. To evaluate our 3D hand pose estimation method, we use the provided 15K/3K split.

RHD is a challenging synthetic dataset that contains 20 subjects performing 39 actions. In total, there are 43,986 rendered hand images of 320×320 resolution, with 41258 images used for training and 2728 for evaluation. For each image, a depth map, a segmentation mask and 2D/3D keypoint annotations are provided. We only used the RGB images and their corresponding 3D labels in our experiments.

We report on two common metrics: 1) Mean End-Point Error (MEPE), which refers to the Euclidean distance between the predicted and the ground truth keypoint locations and 2) the Area Under the percentage of correct keypoints (PCK) Curve (AUC). The distance thresholds of the PCK ranges from 20 mm to 50 mm.

4.2. Implementation Details

We use the PyTorch [28] framework for method implementation. The encoder of the RGB-to-Pose VAE is a ResNet-18 [10], initialized with pretrained weights on the ImageNet dataset [31], whereas the encoder of the Pose-to-Pose VAE and the architecture of the VAE alignment components are similar to [36]. The dimensions of the latent spaces at every stage are set to 128 for the RGB-to-Pose VAE and 64 for the Pose-to-Pose VAE. The hyperparameters $\beta_{VAE}^{cross}, \beta_{VAE}^{uni}, \beta_{align}^{cross}, \beta_{align}^{uni}$ are set to 10^{-5}, the weight w in equation 5 is set to 10^{-2} while r_{kcs} and r_{geo} are set to 10^{-2} and 10^{-6}, respectively. In all training stages we use the Adam optimizer [13] with learning rate 10^{-4} and batch size of 64.

To crop the hand region from the input image, we use the 2D annotations in both datasets to create a bounding box around the hand. To augment data, we consider random rotation in the range $[-45°, 45°]$, random vertical flip with probability 0.5 and image resize to 256×256. Additionally, handedness, palm center and scale of hand are provided during both training and testing. Moreover, we move the center of the hand to the center of the bounding box and accordingly rotate the 3D pose. Therefore the 3D hand pose is aligned with the z-axis of the camera. This process solves the one-to-many mapping, as indicated by [19, 40].

Model	Components/Stages					MEPE
	Injection Decoder	Disent-anglement	Alignment	Refinement KCS	Refinement GEO	
Baseline	×	×	×	×	×	15.71
A	✓	×	×	×	×	15.08
B	✓	✓	×	×	×	14.77
C	✓	×	✓	×	×	14.82
D_0	✓	✓	✓	×	×	14.36
D_1	✓	✓	✓	✓	×	13.93
D_2	✓	✓	✓	×	✓	13.99
D_3	✓	✓	✓	✓	✓	**13.88**

Table 1: Ablation study on the RHD dataset.

Decoder Architecutre	Decoded subspaces	MEPE
Linear Decoder	Concat	15.12
	Pose-specific	15.13
Injection Decoder	Concat	14.87
	Pose-specific	14.77

Table 2: Impact of the injection decoder and decoded subspaces on the disentanglement stage. Concat denotes the concatenated RGB pose-specific and RGB context latent spaces.

4.3. Ablation Study

We evaluate the different components and stages of the proposed method to provide direct insight into their impact on the method's performance (Table 1). We opt to conduct the ablation study on the RHD dataset, since it is the largest of the two datasets and contains heavily occluded fingers, allowing a better demonstration of the performance of the proposed method. In Table 1, Baseline denotes a simple cross-modal RGB-to-Pose VAE. Case A presents the baseline model equipped with the injection decoder. Case B showcases the benefit of the disentanglement stage. Case C

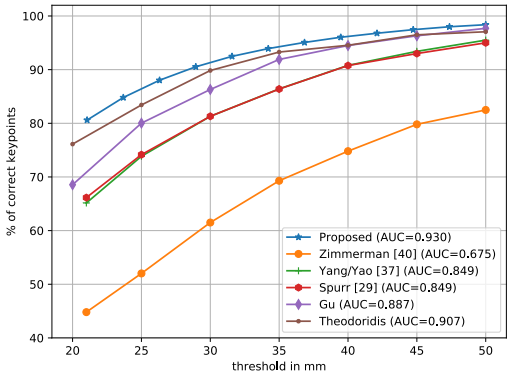
examines the impact of the variational mapping when applied to the model of case A (i.e., without disentanglement). Cases $D_i, i \in 0, ..., 3$, evaluate the impact of the loss terms on the method's performance, applied after the disentanglement and variational alignment stages. Moreover, Table 2 demonstrates the effectiveness of the injection decoder and the decoded subspace in the disentanglement process.

Impact of the injection decoder. To evaluate the benefit of the injection decoder, we compare the MEPE of the baseline model and a model equipped with the injection decoder (case A in Table 1). Table 1 shows that the baseline model has 4% higher relative recognition error as compared to the one with the injection decoder (15.71 versus 15.08 MEPE). This proves the effectiveness of injecting the latent space sample to intermediate decoding layers, thus improving the discrimination ability and decoding accuracy of the constructed latent spaces thanks to the residual connections and the better flow of gradients. This claim is further verified by the results of Table 2 that show the creation of a more descriptive pose-specific latent space during the disentanglement stage.

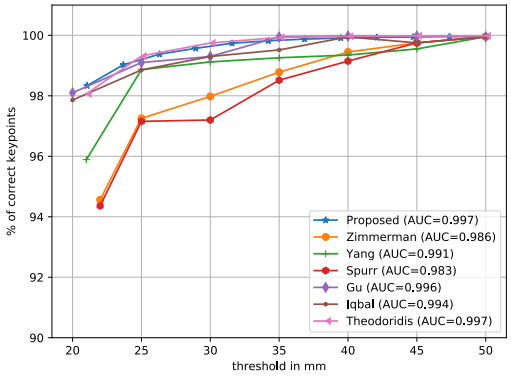
Benefit of the disentanglement stage. Initially, we assess the impact of the injection decoder and the decoded subspaces during the disentanglement stage. The results in Table 2 reveal that the use of the pose-specific latent subspace with the injection decoder outperforms gives 14.77 MEPE, while the use of the concatenated RGB latent subspaces (i.e., concatenation of RGB pose and RGB context latent spaces) results in 14.87 MEPE. On the other hand, the performance of a linear decoder is not affected by the decoded subspace. Therefore, the injection decoder can assist in the construction of a more descriptive pose-specific latent space that can significantly improve the disentanglement stage.

Afterwards, we evaluate the effect of the disentanglement stage when applied on the baseline model (comparison of case A and case B in Table 1). The disentanglement stage improves the hand pose estimation results, leading to lower recognition errors by 2%. Our findings validate the importance of disentangling the RGB context from the pose-specific information, thus managing to significantly reduce the impact of background and illumination variations that are present on the RGB images and improve the hand pose estimation results. Since the injection decoder and the disentanglement training stage are beneficial to the performance of the proposed hand pose estimation methodology, we perform the rest of the experiments without omitting them.

Impact of the variational alignment stage. For the evaluation of the second training step, we employ the proposed alignment components to create two new latent spaces for the pose-specific and the true pose information (case D_0 in Table 1). The results reveal that the variational alignment stage brings the pose-specific information extracted from the RGB images closer to the ground truth pose information, thus improving the hand pose estimation results (MEPE reduction of 2.7%).



(a) RHD



(b) STB

Figure 3: AUC on PCK curve: Comparison against state-of-the-art methods on a) RHD and b) STB datasets.

Evaluation of the refinement stage with the KCS loss. Experiments were conducted to determine the impact of the KCS loss term during the refinement stage (case D_1 in Table 1), from which we observe a significant improvement in the results (3% relative MEPE reduction). This finding validates the importance of applying physiological constraints to the proposed methodology and restricting the predicted 3D hand poses to the space of plausible poses.

Evaluation of the refinement stage with the GEO loss. In a similar fashion with the KCS loss, we evaluate the effect of the GEO loss during the refinement stage (case D_2 in Table 1). The results show that the GEO loss term leads to a reduction in the relative MEPE by 2.5%. This finding demonstrates the importance of applying geometrical constraints that empower the proposed methodology to discard non-plausible poses and finger deformations.

Finally, we combine the KCS and the GEO loss terms

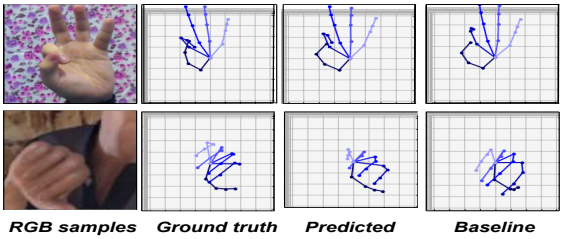


Figure 4: Qualitative results on 3D hand poses from the STB and RHD datasets.

Method	RHD	STB
Zimmerman et al.[47]	30.42	8.68
Moon et al.[23]	20.89	7.95
Yang et al.[41]	19.95	8.66
Spurr et al.[33]	19.73	8.56
Gu et al.[9]	17.11	7.27
Iqbal et al.[12]	15.77	-
Theodoridis et al.[36]	15.61	6.93
Proposed	**13.88**	**6.71**

Table 3: Comparison against state-of-the-art approaches on the RHD and STB datasets

during the refinement stage (case D_3 in Table 1). This is the final proposed hand pose estimation methodology that achieves superior performance with respect to the baseline model (11.6% relative MEPE improvement).

4.4. Comparison with state-of-the-art approaches.

We compare the performance of our proposed method against other state-of-the-art RGB-based Model-free approaches, thus excluding works that process depth images, leverage multimodal input data or employ a hand model [6] for fair comparison. More specifically, the comparative evaluation includes the following approaches: Zimmerman et al. [47], Moon et al.[23], Yang et al.[41], Spurr et al.[33], Gu et al. [9], Iqbal et al. [12] and Theodoridis et al. [36].

EPE comparison. Table 3 summarizes the performance of our proposed hand pose estimation method against other state-of-the-art approaches. For the method of Iqbal et al. [12], we report results only for the RHD dataset (no experiments were performed on the STB dataset) and with the depth maps predicted (ground truth depth maps are not considered for fair comparison with the other approaches). Our proposed multi-stage 3D hand pose estimation methodology outperforms all other methods, yielding 13.88 MEPE on the RHD dataset and 6.71 MEPE on the STB dataset.

PCK comparison. We compare the PCK curves of our method and the other state-of-the-art approaches on the RHD and STB datasets and report the results in Figure 3. On the RHD dataset, our method achieves an AUC score of 0.930, clearly outperforming all other state-of-the-art methods, while on the STB dataset, our method achieves an AUC score of 0.997, which is on par with the method of [36].

From the experimental results, we can observe that the proposed methodology surpasses all state-of-the-art methods on both datasets using MEPE and PCK metrics. More importantly, the overall performance improvement is greater on the RHD dataset, despite the fact that the RHD dataset is large and challenging due to a wide number of self-occluded fingers, backgrounds and subjects. On the other hand, the STB dataset is considerably smaller and saturated as it contains a single subject's left hand within a limited number of different backgrounds. As a result, the STB dataset is not optimal for thoroughly demonstrating the full capabilities of the proposed methodology.

Qualitative results comparison. Finally, we perform a qualitative evaluation of the poses predicted by our method. Figure 4 illustrates several predicted poses of the proposed multi-stage 3D hand pose estimation method, compared to the baseline RGB-to-Pose VAE and the ground truth 3D poses. The proposed model predicts 3D hand joint locations with higher precision in both the small and saturated STB dataset, as well as the large and challenging RHD dataset, demonstrating its ability to overcome the difficulties imposed by occlusions, background and illumination variations and different camera positions.

5. Conclusions

This paper presents a novel multi-stage RGB-based approach for accurate 3D hand pose estimation. To this end, a GAN is initially used to disentangle the pose-specific information of the RGB images from the irrelevant RGB context. Subsequently, two variational mappers project the pose-specific and ground truth pose latent spaces to new latent spaces that are better aligned with each other. These operations ensure an optimal cross-modal alignment between the RGB and 3D pose information. Finally, two loss terms are employed to ensure that the predicted poses abide by kinematic and geometrical constraints, thus avoiding non-plausible poses. A novel injection decoder is also proposed to improve the discrimination ability and decoding accuracy of the constructed latent spaces. A thorough ablation study and extensive experimental results on two well-known datasets demonstrate the benefits of each component and stage of the proposed method, as well as the method's superiority against other state-of-the-art approaches.

6. Acknowledgement

This work was supported by the Greek General Secretariat of Research and Technology under contract T1EΔK-02469 EPIKOINONO.

References

[1] M. Abavisani, H. R. V. Joze, and V. M. Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1165–1174, 2019. 1

[2] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020. 1

[3] M. Alivizatou-Barakou, A. Kitsikidis, F. Tsalakanidou, K. Dimitropoulos, C. Giannis, S. Nikolopoulos, S. Al Kork, B. Denby, L. Buchman, M. Adda-Decker, et al. Intangible cultural heritage and new technologies: challenges and opportunities for cultural preservation and development. In *Mixed reality and gamification for cultural heritage*, pages 129–158. Springer, 2017. 1

[4] S. Baek, K. I. Kim, and T.-K. Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2

[5] G. Caggianese, N. Capece, U. Erra, L. Gallo, and M. Rinaldi. Freehand-steering locomotion techniques for immersive virtual environments: A comparative evaluation. *International Journal of Human–Computer Interaction*, 36(18):1734–1755, 2020. 1

[6] T. Chatzis, A. Stergioulas, D. Konstantinidis, K. Dimitropoulos, and P. Daras. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 10(19):6850, 2020. 2, 8

[7] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10833–10842, 2019. 2

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[9] J. Gu, Z. Wang, W. Ouyang, J. Li, L. Zhuo, et al. 3d hand pose estimation with disentangled cross-modal latent space. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 391–400, 2020. 1, 2, 8

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6

[11] Y. He, W. Hu, S. Yang, X. Qu, P. Wan, and Z. Guo. 3d hand pose estimation in the wild via graph refinement under adversarial learning. *arXiv*, pages arXiv–1912, 2019. 2

[12] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 1, 2, 8

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3

[15] D. Konstantinidis, K. Dimitropoulos, and P. Daras. Sign language recognition based on hand and body skeletal data. In *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2018. 1

[16] D. Konstantinidis, K. Dimitropoulos, K. Stefanidis, T. Kalvourtzis, S. Gannoum, N. Kaklanis, K. Votis, P. Daras, S. Rovira-Esteva, P. Orero, et al. Developing accessibility multimedia services: the case of easytv. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–8, 2020. 1

[17] O. Kopuklu, N. Kose, and G. Rigoll. Motion fused frames: Data level fusion strategy for hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2103–2111, 2018. 1

[18] T. Lee and T. Hollerer. Multithreaded hybrid feature tracking for markerless augmented reality. *IEEE transactions on visualization and computer graphics*, 15(3):355–368, 2009. 1

[19] S. Li and D. Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11927–11936, 2019. 6

[20] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7113–7122, 2020. 2

[21] A. Markussen, M. R. Jakobsen, and K. Hornbæk. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1073–1082, 2014. 1

[22] G. Moon, J. Yong Chang, and K. Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 2

[23] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *arXiv preprint arXiv:2008.09309*, 2020. 2, 8

[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[25] M. Oberweger and V. Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 585–594, 2017. 1

[26] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 1

[27] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras. Continuous sign language recognition through

cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, 8:91170–91180, 2020. 1

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 6

[29] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *IFIP Conference on Human-Computer Interaction*, pages 282–299. Springer, 2013. 1

[30] P. Ren, H. Sun, Q. Qi, J. Wang, and W. Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 2

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[32] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. *arXiv preprint arXiv:2003.09282*, 2020. 1, 2

[33] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 1, 2, 8

[34] S. Sridhar, A. M. Feit, C. Theobalt, and A. Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3643–3652, 2015. 1

[35] K. Stefanidis, D. Konstantinidis, A. Kalvourtzis, K. Dimitropoulos, and P. Daras. 3d technologies and applications in sign language. *Recent Advances in 3D Imaging, Modeling, and Reconstruction*, pages 50–78, 2020. 1

[36] T. Theodoridis, T. Chatzis, V. Solachidis, K. Dimitropoulos, and P. Daras. Cross-modal variational alignment of latent spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 960–961, 2020. 1, 2, 6, 8

[37] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014. 1

[38] B. Wandt, H. Ackermann, and B. Rosenhahn. A kinematic chain space for monocular motion capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 5

[39] B. Wandt and B. Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7782–7791, 2019. 2, 5

[40] L. Yang, S. Li, D. Lee, and A. Yao. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2335–2343, 2019. 2, 6

[41] L. Yang and A. Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 1, 2, 8

[42] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 2, 6

[43] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157. IEEE, 2017. 5

[44] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2354–2364, 2019. 2

[45] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. In *IJCAI*, 2016. 1

[46] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 2

[47] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 1, 2, 6, 8