SEAL-Pose: Enhancing 3D Human Pose Estimation through Trainable Loss Function

Junggeun Do*
Texas A&M University
jg.do@tamu.edu

Jay-Yoon Lee
Seoul National University
lee.jayyoon@snu.ac.kr

Abstract

We propose SEAL-Pose, a method that trains models to predict more plausible 3D human poses through a trainable loss function that dynamically learns the output structures of data. While inspired by the idea of Structured Energy As Loss (SEAL), initially designed for structured prediction and limited to probabilistic models with relatively simple output structures, we extend it to tackle 3D human pose estimation, a task with more complex and high-dimensional structural dependencies than those considered in previous applications. SEAL-Pose enables pose estimation models (task-net) to learn joint dependencies via trainable loss function (loss-net) that automatically capture body structure during training without explicit prior knowledge and is applicable to any backbone models. We also suggest evaluation metrics such as the limb symmetry error (LSE) and body segment length error (BSLE) to assess the structural consistency of the predicted poses. These metrics measure overall structural preservation, which the vast majority of existing metrics do not capture. Experimental results on the Human3.6M, MPI-INF-3DHP, and Human3.6M WholeBody datasets show that SEAL-Pose not only reduces per-joint pose estimation errors but also generates more plausible poses. In addition, SEAL-Pose demonstrates more significant improvements in challenging settings such as monocular single-frame pose estimation. Our work also highlights the potential of employing trainable loss functions for tasks with complex output structures, offering a promising direction for future research.

1. Introduction

Pose estimation is a critical task in computer vision that requires accurate prediction of keypoint positions of objects or humans. In particular, 3D human pose estimation (3D HPE) is even more challenging because it involves predicting spatial structures while adhering to anatomical constraints [12]. However, common training objectives such as mean squared error (MSE) and mean per-joint position error (MPJPE) penalize individual joint errors without accounting for structural consistency, which often results in implausible or anatomically inconsistent poses. To address these issues, it is critical to effectively model the dependencies in the output space to predict accurate and plausible 3D poses. There have been previous studies [3, 8, 22, 23, 27] that attempted to capture the structural dependencies of human poses, but they often rely on manually designed rules or specific model architectures, which constrain their adaptability and scalability.

To overcome these limitations, we propose SEAL-Pose, a novel framework that employs a trainable loss function to provide structural guidance for 3D pose estimation without requiring explicit priors. At the core of this framework is the loss-net, a neural network jointly optimized with the pose estimation model (task-net). The loss-net learns to capture dependencies among joints and dynamically evaluates pose plausibility during training, unlike conventional per-joint error objectives. Building on the Structured Energy As Loss (SEAL) framework [11], initially applied to generic multi-label classification problems and natural language applications, we extend the concept of trainable loss functions to the 3D HPE.

Our proposed method enables pose estimation models to learn joint dependencies during training, allowing the model to more accurately represent relationships in the output space. This results in improved 3D HPE performance in terms of widely used per-joint error metrics, such as mean per-joint position error (MPJPE). Unlike previous methods that manually encode body structures or use domain-specific rules [3, 8, 22, 23, 27], SEAL-Pose automatically captures joint dependencies without requiring predefined structural priors, through a trainable loss function. This approach offers a flexible and scalable solution that is applicable to any backbone models.

Additionally, we also suggest new evaluation metrics, such as Limb Symmetry Error (LSE) and Body Segment Error (BSLE), to evaluate the structural consistency of pre-

*Work performed while at Seoul National University.

dicted poses. Our experimental results on the Human3.6M, MPI-INF-3DHP, and Human3.6M WholeBody datasets demonstrate that SEAL-Pose not only reduces per-joint errors but also produces more anatomically plausible poses, evaluated by LSE and BSLE. SEAL-Pose shows even more improvements in challenging settings like monocular single-frame pose estimation, highlighting its potential for broader applications. Overall, empirical results suggest that our approach could be applied to a wide range of tasks that require capturing complex dependencies in the output space.

2. Related Work

2.1. 3D Human Pose Estimation

3D human pose estimation is a well-established computer vision task involving the prediction of 3D joint positions from 2D images or videos. This task is inherently challenging because it requires inferring spatial relationships and ensuring anatomical plausibility using limited visual information. Current 3D HPE approaches typically follow two paradigms: (1) directly predicting 3D poses from images [17, 18] or (2) using a two-step process where 2D poses are estimated first and then lifted to 3D space [12, 26]. The latter approach has become more popular and effective following advances in 2D human pose estimation [26]. Therefore, we also adopted the 2D-to-3D lifting approach in our work.

3D whole-body pose estimation extends traditional 3D human pose estimation by integrating detailed annotations for additional keypoints, including those for the face, hands, and feet, demanding more fine-grained and precise predictions. The expanded scope introduces greater challenges due to the variation in scales and the increased diversity of poses associated with detail keypoints. Recently, [28] developed the Human3.6M 3D WholeBody dataset (H3WB) based on the widely used Human3.6M dataset (H36M) by including annotations for 133 keypoints, as shown in Figure 1. This dataset has become an important resource, allowing research to address the increased complexity of whole-body pose estimation while encouraging methods that go beyond traditional approaches focused mainly on standard body keypoints.

2.2. Output Structure of 3D HPE

2D-to-3D HPE has inherent challenges such as ambiguity due to incomplete information, which is further compounded in single-frame scenarios. To address this issue, several works have focused on capturing the structural dependencies between body joints. For instance, [27] proposed the Joint Relationship Aware Network, which enhances pose predictions by considering both global and local joint relationships. [22] introduced the Limb Poses Aware Network, which incorporates relative and absolute bone angles to model pose structure. However, these methods tend to be closely tied to specific model architectures. Another notable approach is Pose Grammar [3, 23], which uses predefined kinematic rules and bidirectional recurrent neural networks to refine pose predictions.

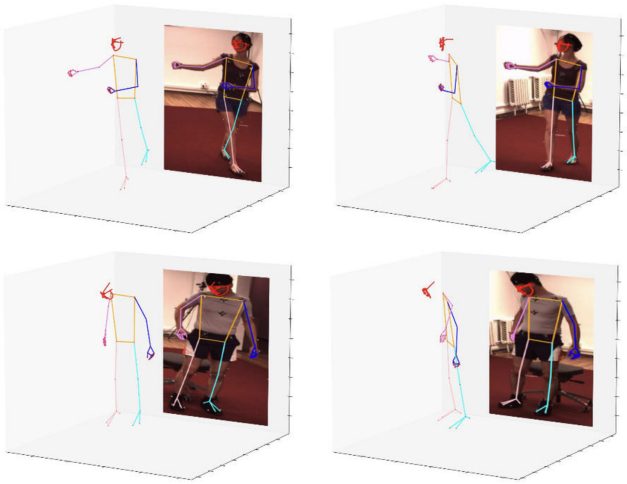

Figure 1. **Example Annotations from the H3WB Dataset**. H3WB extends Human3.6M with keypoints for hands, face, and feet, enabling detailed whole-body pose estimation.

More recently, some works have proposed methods using multiple hypotheses or generating plausible 3D poses to overcome the inherent difficulties of 3D HPE. For example, [8] proposed Biomechanical Pose Generator to augment training data with biomechanically plausible poses. They also introduced Binary Depth Coordinates to resolve the depth ambiguity by classifying the joint depths as front or back. Additionally, [20] suggested ManiPose, a manifold-constrained multi-hypothesis approach to overcome depth ambiguity through estimating the plausibility of each hypothesis and constraining them to the human pose manifold.

Despite their contributions, most previous methods rely on expert knowledge or predefined rules to capture joint dependencies, which may limit their scalability and adaptability. In contrast, we aim to address these limitations by providing a more flexible and scalable approach for 3D HPE that captures joint dependencies without explicit prior knowledge. Our method, SEAL-Pose, can be applied to any backbone models as well as potentially extending to various tasks with complex output structures.

2.3. Structured Energy As Loss (SEAL)

SEAL [11] builds on the concept of using structured energy networks for structured prediction, initially introduced by [1]. These early models, known as Structured Prediction Energy Networks (SPENs), effectively captured arbitrary global dependencies in the output space without explicitly representing them. However, they were limited to slow and unstable inference due to the inherent problem of updating

output variables through gradient-based inference (GBI).

SEAL addresses this issue by using structured energy networks as trainable loss functions rather than direct predictors, leveraging the expressivity of energy networks while enabling faster and more stable inference at test time. SEAL has been applied to tasks such as multi-label classification, semantic role labeling, and image segmentation, highlighting its potential to improve performance and efficiency over traditional methods. However, SEAL can only be applied to probabilistic models because it utilizes the output distribution of a neural network, referred to as a task-net, as a dynamic noise distribution to train a structured energy network, referred to as a loss-net.

Specifically, SEAL is implemented in two main ways: SEAL-static and SEAL-dynamic. SEAL-static uses a pre-trained, fixed loss-net to guide the task-net, while SEAL-dynamic updates the loss-net dynamically based on the evolving outputs of the task-net. SEAL-dynamic could be more beneficial for task-net training since it can better focus on the current distribution of task-net predictions. [11] has shown that SEAL-dynamic generally outperforms SEAL-static across structured prediction tasks, as it captures dependencies more effectively and provides more helpful learning signals even in high-dimensional spaces.

Our work expands upon these foundations by introducing a novel application of the SEAL framework. We applied SEAL to deterministic models, particularly for 3D HPE in 2D-to-3D lifting scenarios. Since it is not possible to extract negative samples from the task-net's output distribution because deterministic models directly output real-valued predictions, we used the task-net's predictions themselves as negative samples for the loss-net. This intuitive approach has been shown to be effective in practice, as long as the batch size is sufficiently larger than the task-net, allowing the loss-net to learn meaningful structural dependencies during training.

Unlike existing pose estimation methods, in this way, SEAL-Pose offers a novel approach to capturing joint dependencies in the output space during training, even without any explicit prior knowledge, through a trainable loss function. SEAL-Pose allows for more accurate and coherent 3D pose predictions and offers a flexible and scalable method to improve pose estimation. We also seek the potential of using a trainable loss function to model dependencies in the output space to improve various tasks with complex output structures, which is a promising direction for further research.

3. Methodology

3.1. SEAL-Pose

SEAL-Pose extends the SEAL framework for 3D human pose estimation, particularly in a 2D-to-3D lifting scenario. SEAL-Pose provides a trainable loss function that automatically captures joint dependencies as the model trains, without manually encoded body structure and predefined rules, unlike previous approaches [3, 8, 22, 23, 27]. By incorporating the SEAL framework, our method allows pose estimation models to better capture the relationships between joints, leading to more accurate and coherent 3D pose predictions, utilizing the capacity of structured energy networks to well model the dependencies in output space. Moreover, SEAL-Pose is applicable to any model architecture and can be combined with any data augmentation method, offering flexibility and scalability.

In particular, we implement the SEAL-dynamic approach to take advantage of the superiority of SEAL-dynamic over SEAL-static, as presented in [11]. Therefore, in SEAL-Pose, the pose estimation model (task-net)[1] and the structured energy network (loss-net)[1] are trained jointly, by dynamically updating the loss-net based on the current predictions of the task-net. This iterative joint optimization process ensures that the loss-net remains synchronized with the task-net's progress, enhancing its ability to guide the task-net effectively. This approach leads to more accurate and structurally consistent 3D pose predictions by dynamically modeling joint dependencies during training.

In this framework, the task-net $F_\phi(x)$ is optimized to minimize a weighted sum of the mean squared error (MSE) loss and the output of the the loss-net (energy) $E_\theta(x, \tilde{y})$. Specifically, the task-net parameters ϕ are updated using the following manner:

$$\phi_t \leftarrow \phi_{t-1} - \eta_\phi \nabla_\phi \frac{1}{|B_t|} \sum_{(x,y)\in B_t} L_F(\phi; \theta) \tag{1}$$

where B_t is the mini-batch of training samples at iteration t, η_ϕ is the learning rate for the task-net, and $L_F(\phi; \theta)$ is the combined loss function. The combined loss function is defined as:

$$L_F(x_i, y_i; \theta) = \sum_{j=1}^{M} \text{MSE}(y_j, F_\phi(x)_j) + \alpha E_\theta(x, F_\phi(x)) \tag{2}$$

where M refers to the total number of joints in the pose estimation dataset and x represents the input data, specifically the 2D joint coordinates. The variable y_j denotes the ground-truth 3D joint coordinates, while $F_\phi(x)_j = \tilde{y}_j$ represents the predicted 3D joint coordinates from the task-net. The energy term $E_\theta(x, F_\phi(x))$ is computed by the loss-net and implicitly evaluates the structural dependencies between joints. Finally, α is a hyperparameter controlling the balance between the MSE loss and the energy term.

[1] In the rest of this section, we refer to the pose estimation model as task-net and the structured energy network as a trainable loss function as loss-net.

Algorithm 1 SEAL-Pose Algorithm

Require: T: number of training iterations
Require: $(\mathbf{x}, \mathbf{y})$: training data (2D inputs and 3D labels)
Require: F_ϕ: task-net with parameters ϕ
Require: E_θ: loss-net with parameters θ
Require: $\texttt{optimizer}_\phi$: optimizer for task-net
Require: $\texttt{optimizer}_\theta$: optimizer for loss-net
  Initialize ϕ_0, θ_0 randomly
  for $t = 1$ to T **do**
    Sample mini-batch $B_t = \{(x_i, y_i)\}_{i=1}^N$ from training data
    Compute task-net outputs: $\tilde{y}_i = F_{\phi_{t-1}}(x_i)$ for all $x_i \in B_t$
    Update loss-net parameters θ_t:
      $\theta_t \leftarrow \theta_{t-1} - \eta_\theta \nabla_\theta \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} L_E(x_i, y_i, \tilde{y}_i\, \theta)$
    Update task-net parameters ϕ_t:
      $\phi_t \leftarrow \phi_{t-1} - \eta_\phi \nabla_\phi \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} L_F(x_i, y_i; \theta_t)$
  end for

The loss-net is dynamically trained to adapt to the task-net's predictions by minimizing the loss L_E:

$$\theta_t \leftarrow \theta_{t-1} - \eta_\theta \nabla_\theta \frac{1}{|B_t|} \sum_{(x,y)\in B_t} L_E(x, y, F_{\phi_{t-1}}(x); \theta) \tag{3}$$

We employ two types of loss for L_E: margin-based loss and a simplified form of noise contrastive estimation (NCE) ranking loss [13], both suggested in [11].

The margin-based loss enforces the loss-net to decrease the energy $E_\theta(x, y)$ of the ground truth label y and increase the energy $E_\theta(x, \tilde{y})$ of the task-net's incorrect prediction $\tilde{y}$, such that the difference between the two energies is sufficiently large to exceed the margin. The margin-based loss is defined as:

$$L_E^{\text{margin}}(x_i, y_i, \tilde{y}_i; \theta) = \max_{\tilde{y}} [\Delta(y, \tilde{y}) - E_\theta(x, \tilde{y}) + E_\theta(x, y)]_+ \tag{4}$$

where $\Delta(y, \tilde{y})$ denotes a task-specific margin function, MPJPE in our implementation.

Similarly, the NCE ranking loss minimizes the energy of the ground truth label y while increasing the energy of the task-net's prediction $\tilde{y}$, treating the task-net's predictions as negative samples. The NCE ranking loss is defined as:

$$L_E^{\text{NCE}}(x_i, y_i, \tilde{y}_i; \theta) = -\log \frac{\exp(-E_\theta(x, y))}{\exp(-E_\theta(x, y)) + \exp(-E_\theta(x, \tilde{y}))} \tag{5}$$

Additionally, we used a larger mini-batch, which always includes the entire mini-batch for the task-net, in updating the loss-net in order to improve loss-net training.

In SEAL-Pose, the task-net and loss-net are updated in an alternative manner, enabling the loss-net to adapt dynamically to the task-net thus improving 3D pose predictions by teaching dependencies in the output space to task-net more effectively. This iterative joint optimization process is summarized in Algorithm 1.

Algorithm 2 Gradient-Based Inference

Require: $(\mathbf{x}, \mathbf{y})$: training data (2D inputs and 3D labels)
Require: F_ϕ: task-net, E_θ: energy network
Require: T: training iterations, K: GBI steps
  Phase 1: train task-net
  for $t = 1$ to T **do**
    Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
    Update ϕ:
      $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} \text{MSE}(F_\phi(x_i) - y_i)$
  end for
  Phase 2: train energy network
  for $t = 1$ to T **do**
    Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
    Generate $\tilde{y}_i = F_\phi(x_i)$ for $x_i \in B_t$
    Update θ:
      $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} [E_\theta(x_i, y_i) - E_\theta(x_i, \tilde{y}_i)]$
  end for
  Phase 3: gradient-based inference
  Initialize $\tilde{y}_i^{(0)} = F_\phi(x_i)$ for $x_i \in B_t$
  for $k = 1$ to K **do**
    Refine $\tilde{y}_i$: $\tilde{y}_i^{(k)} \leftarrow \tilde{y}_i^{(k-1)} - \eta \nabla_{\tilde{y}} E_\theta(x_i, \tilde{y}_i^{(k-1)})$
  end for

3.2. Gradient-Based Inference

We implement a gradient-based inference (GBI) method with trained loss-net and pose-net, to examine whether the loss-net effectively captures structural dependencies in human poses. GBI is a method that leverages gradients to iteratively refine the outputs [1, 4–6, 16] or parameters [10] of neural networks, and we adopt the former approach. Specifically, we iteratively update the predictions of pose-net along the gradient provided by the loss-net, with the objective of decreasing the energy. This procedure provides a direct way to evaluate whether the learned energy function captures human pose structure.

We used GBI in two complementary ways. First, we compared the effectiveness of the structured energy network when used as a trainable loss function versus as a direct predictor, providing insights into whether incorporating it as loss-net yields more consistent improvements. The detailed GBI procedure for this comparison is summarized in Algorithm 2. Second, we employed GBI as an analysis tool to directly probe the gradient signals of the loss-net trained jointly with the task-net in SEAL-Pose. In this case, if the loss-net has successfully captured structural dependencies in human poses, then following its gradients should iteratively refine task-net predictions toward more plausible poses, offering a direct way to examine the quality of the learned energy function.

4. Experiments

4.1. Datasets

We conduct our empirical experiments on Human3.6M dataset (H36M) [7], MPI-INF-3DHP (3DHP) [15] dataset and Human3.6M 3D WholeBody dataset (H3WB) [28]. H36M is the most widely used dataset for 3D human pose estimation [12, 26]. 3DHP is a more challenging dataset than H36M because it contains fewer samples and includes both indoor and outdoor scenes, while H36M only contains indoor scenes. H3WB is a recent dataset for 3D whole-body pose estimation. H3WB extends H36M by providing whole-body keypoint annotations with detailed information about hands, face, and feet, making it suitable for evaluating fine-grained 3D pose estimation. We utilize the ground truth 2D joint coordinates provided in the datasets to align the 3D and 2D poses. For the H36M and 3DHP datasets, we zero-center the 3D poses around the pelvis joint, following prior works. For the H3WB dataset, we zero-center the 3D poses around the midpoint of the two hip joints, following the dataset's protocol.

4.2. Implementation Details

There are two main settings for 3D human pose estimation: single-frame and multi-frame pose estimation. In the single-frame setting, models predict 3D poses from a single frame of 2D keypoints, while models exploit richer spatial and temporal information from multiple frames to estimate 3D poses in the multi-frame setting. We evaluate SEAL-Pose mainly on single-frame models to verify its effectiveness in more challenging scenarios. Moreover, we also applied it to multi-frame models to verify the applicability of SEAL-Pose to state-of-the-art models. Specifically, we employed the SimpleBaseline [14], SemGCN [25] and VideoPose [19] as task-net for single-frame pose estimation. For multi-frame pose estimation, we used the MixSTE (input sequence length T=243) [24] and P-STMO (T=81) [21] as task-net for H36M and 3DHP each.

We have modified the input and output layers of task-net to align with the dimensions of each dataset. For the loss-net, we additionally adjusted the SimpleBaseline architecture by modifying the dimensions and depth of the hidden layers. We set the hidden size to 2048 with 2 residual block stages and omitted batch normalization and dropout layers. We used separate Adam optimizers [9] without learning rate decay for the loss-net and task-net. All single-frame models are trained with a batch size of 1024 for 50 epochs on H36M and 3DHP, and a batch size of 64 for 200 epochs on H3WB, and we used reported hyperparameters in their original papers for multi-frame models.

5. Evaluation Metrics

We evaluate our models using standard metrics for 3D human pose estimation. For the H36M dataset, we report MPJPE and P-MPJPE (procrustes-aligned MPJPE), following established protocols [7]. MPJPE measures the average Euclidean distance between the predicted and ground truth 3D joint positions, and P-MPJPE is a more robust metric that considers the alignment of the predicted poses. For the 3DHP dataset, we report MPJPE, PCK (percentage of correct keypoints) within a 150 mm range, and AUC (area under the curve) as evaluation metrics following previous works [15, 21, 24]. On the H3WB dataset, we use the official benchmark's PA-MPJPE (pelvis-aligned MPJPE), measuring per-joint errors for the whole body, body, hands, face, wrist-aligned hands, and nose-aligned face. All reported metrics are averaged over the entire test set of each dataset.

To further assess structural consistency in the predicted poses, we suggest two additional metrics: Limb Symmetry Error (LSE) and Body Segment Length Error (BSLE). These metrics evaluate the structural plausibility of predicted poses by measuring the symmetry between left and right limbs and the predicted lengths of body segments, respectively. Lower LSE and BSLE values indicate more anatomically plausible poses. We provide detailed definitions of LSE and BSLE in the following sections.

5.1. Limb Symmetry Error (LSE)

The Limb Symmetry Error evaluates left-right body symmetry by comparing the lengths of corresponding limbs on the left and right sides. It is defined as the normalized difference in lengths between each pair of corresponding limbs, such as wrist-to-elbow and ankle-to-knee.

Given a set of n corresponding limb pairs, where the i-th left limb is defined by keypoints $\mathbf{l}_{i1}, \mathbf{l}_{i2}$ and the corresponding right limb by $\mathbf{r}_{i1}, \mathbf{r}_{i2}$, the LSE for limb pair i is computed as:

$$\text{LSE}_i = 100 \cdot \left| \frac{\|\mathbf{l}_{i1} - \mathbf{l}_{i2}\| - \|\mathbf{r}_{i1} - \mathbf{r}_{i2}\|}{(\|\mathbf{l}_{i1} - \mathbf{l}_{i2}\| + \|\mathbf{r}_{i1} - \mathbf{r}_{i2}\|)/2} \right|$$

where $\|\cdot\|$ denotes the Euclidean norm.

This metric is calculated based on the predicted poses and measures the relative difference in lengths between the left and right limbs. A value of 0 indicates perfect symmetry, and a larger value indicates worse symmetry, which is desirable for anatomically plausible poses. If the length of the left and right limbs differs $n\%$ from their average length, the LSE value is computed as n.

5.2. Body Segment Length Error (BSLE)

The Body Segment Length Error measures deviations in the lengths of body segments—pairs of adjacent joints—by

Table 1. **Performances on the Human3.6M Dataset**. SEAL-Pose improves MPJPE and P-MPJPE across models.

Metric	MPJPE ↓	P-MPJPE ↓
SimpleBaseline	43.8	34.7
+ SEAL-Pose (margin)	**42.5**	33.9
+ SEAL-Pose (NCE)	42.7	**33.8**
SemGCN	47.0	37.9
+ SEAL-Pose (margin)	45.1	36.6
+ SEAL-Pose (NCE)	**44.9**	**36.5**
VideoPose	41.6	32.4
+ SEAL-Pose (margin)	**41.0**	**32.3**
+ SEAL-Pose (NCE)	41.3	32.5
MixSTE (T=243)	20.1	15.7
+ SEAL-Pose (margin)	**20.0**	**15.5**
+ SEAL-Pose (NCE)	**20.0**	15.6

Table 2. **Performances on the MPI-INF-3DHP Dataset**. SEAL-Pose consistently reduces MPJPE and improves PCK and AUC.

Metric	MPJPE ↓	PCK ↑	AUC↑
SimpleBaseline	80.9	86.9	53.8
+ SEAL-Pose (margin)	**71.8**	**89.3**	**58.7**
+ SEAL-Pose (NCE)	72.3	89.2	58.2
SemGCN	74.5	89.5	56.4
+ SEAL-Pose (margin)	**71.8**	**90.4**	**57.9**
+ SEAL-Pose (NCE)	72.5	90.1	57.7
VideoPose	66.4	90.8	60.5
+ SEAL-Pose (margin)	64.1	91.4	**62.1**
+ SEAL-Pose (NCE)	**64.0**	**91.7**	**62.1**
P-STMO (T=81)	32.8	**98.3**	77.7
+ SEAL-Pose (margin)	**32.2**	98.2	**78.1**
+ SEAL-Pose (NCE)	32.4	98.1	**78.1**

comparing predicted poses and ground-truth poses. For each segment i, with predicted adjacent keypoints $\mathbf{k}_{i1}, \mathbf{k}_{i2}$ and corresponding ground truth keypoints $\mathbf{t}_{i1}, \mathbf{t}_{i2}$, BSLE is defined as:

$$\text{BSLE}_i = 100 \cdot \left| 1 - \frac{\|\mathbf{k}_{i_2} - \mathbf{k}_{i_1}\|}{\|\mathbf{t}_{i_2} - \mathbf{t}_{i_1}\|} \right|$$

We refer to the specific case of BSLE that focuses only on limb segments as the limb length error (LLE). These metrics calculate the relative difference between predicted and ground truth segment lengths, reflecting how well the model preserves anatomical proportions. Lower BSLE and LLE values indicate better preservation of segment lengths in the predicted poses. If the length of the predicted segment differs $n\%$ from the ground truth segment length, the BSLE value is computed as n.

6. Experimental Results and Analysis

6.1. Pose Estimation Error Evaluation

Human3.6M Dataset We evaluated the impact of SEAL-Pose on various models on the H36M dataset. As shown in Table 1, SEAL-Pose consistently outperformed the baseline models across all metrics, more notably with the single-frame models. These results demonstrate that SEAL-Pose effectively reduces 3D pose estimation errors, especially in more challenging settings.

MPI-INF-3DHP Dataset For the 3DHP dataset, SEAL-Pose also demonstrated consistent improvements across all models, as shown in Table 2. SEAL-Pose showed a larger performance gap on the 3DHP dataset, which is more challenging than H36M due to its diverse scenes and fewer samples, where the model requires more guidance to capture complex dependencies and structures. Moreover, SEAL-Pose provided more substantial improvements on the SimpleBaseline model, which has a more straightforward architecture compared to the others. These results, which show that SEAL-Pose is more beneficial in difficult settings, clearly suggest the strength of SEAL-Pose in providing structural awareness that is not sufficient for MSE loss. For the P-STMO task-net, SEAL-Pose showed a slight decrease in PCK but achieved improvements in MPJPE and AUC, which could outweigh the PCK loss, implying that it can still be advantageous for state-of-the-art models.

Table 3. **Performance on the Human3.6M WholeBody Dataset**. SEAL-Pose reduces pelvis-aligned MPJPE across all body parts, resulting in more coherent predictions. † from H3WB's official benchmark. ‡ nose-aligned MPJPE for face and wrist-aligned MPJPE for hands.

Method	Whole-body	Body	Face/Aligned‡	Hand/Aligned‡
Jointformer †	88.3	84.9	66.5 / 17.8	125.3 / 43.7
3D-LFM (Dabhi et al. [2])	64.1	60.8	56.6 / 10.4	78.2 / 28.2
SimpleBaseline	65.5	62.8	49.6 / 14.6	92.7 / 35.1
+ GBI	65.3	62.6	49.4 / 14.8	92.5 / 35.0
+ SEAL-Pose (margin)	**62.8**	**61.1**	**46.3** / **13.7**	**90.7** / 34.7
+ SEAL-Pose (NCE)	63.4	**61.1**	46.5 / 14.5	92.1 / **34.2**
VideoPose	60.1	56.4	46.3 / 11.9	84.3 / 29.6
+ GBI	60.0	56.3	46.3 / 12.4	84.2 / 29.5
+ SEAL-Pose (margin)	**58.6**	55.7	**45.0** / 11.6	**82.3** / 29.3
+ SEAL-Pose (NCE)	58.8	**54.8**	45.5 / **11.5**	82.7 / **28.9**

Human3.6M WholeBody Dataset To evaluate the impact of SEAL-Pose on 3D whole-body pose estimation with detailed annotations and complex body structures, we conducted experiments on the H3WB dataset. As shown in Table 3, SEAL-Pose consistently outperformed the baseline models across all body parts, reducing the pelvis-aligned MPJPE, demonstrating its effectiveness in improving 3D whole-body pose estimation and the loss-net's capacity to

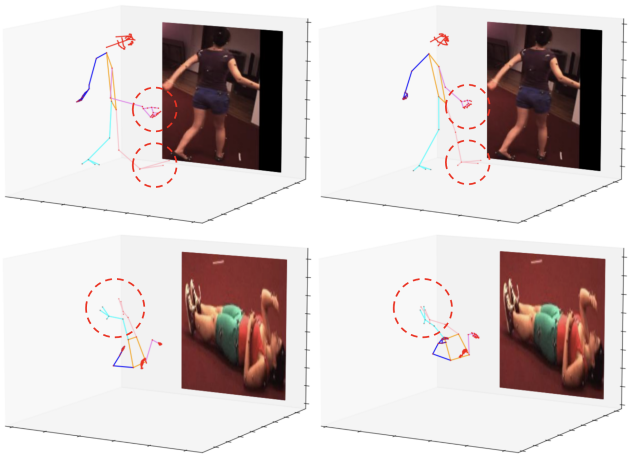

Figure 2. **Comparison of Predicted Poses on H3WB**. Predictions from SimpleBaseline (left) and the same model with SEAL-Pose (right) show improved accuracy, especially in challenging poses. Key differences are highlighted with red circles.

capture complex human body structures, including finer anatomical details like facial features and hand articulations. The improved performance validates SEAL-Pose's ability to model intricate interdependencies among body regions for more accurate and cohesive predictions. Notably, SEAL-Pose showed relatively better performance than the baseline on less common data distributions, such as target figures in reverse and lying down, as illustrated in Figure 2. The superior performance of SEAL-Pose on H3WB highlights its potential to improve 3D human pose estimation, particularly in challenging whole-body settings. Overall, experimental results indicate that SEAL-Pose can provide informative signals to enhance the task-net's learning. In addition, it is particularly effective in challenging settings, requiring more guidance to capture complex dependencies and structures.

Comparison with Gradient-Based Inference In order to compare the effectiveness of SEAL-Pose with GBI, we evaluated the performance improvement of the task-net's predictions using GBI on H3WB. As shown in Table 3, SEAL-Pose consistently outperformed GBI, and the benefits of GBI were limited in our experiments. This result demonstrates that using the structured energy network as a trainable loss function is more effective than using it as a direct predictor, as it better captures dependencies among output variables and teaches the task-net to generate more accurate and coherent 3D pose predictions.

6.2. Pose Structure Evaluation

We evaluated structural consistency by examining the LSE, LLE, and BSLE metrics on the H36M, 3DHP, and H3WB datasets. SEAL-Pose consistently showed lower error values across all three structural metrics on H36M and 3DHP datasets, as detailed in Table 4. These results indicate that SEAL-Pose effectively captures structured dependencies in human poses, leading to more anatomically plausible and consistent 3D pose predictions.

Table 4. **Structural Consistency Evaluation Across Datasets**. SEAL-Pose reduces LSE, LLE, and BSLE, improving plausibility.

Dataset	Metric	LSE ↓	LLE ↓	BSLE ↓
H36M	Ground Truth	0.00	0.00	0.00
	SimpleBaseline	4.85	5.09	6.12
	+ SEAL-Pose (margin)	4.72	**4.61**	**5.46**
	+ SEAL-Pose (NCE)	**4.68**	4.65	5.70
3DHP	Ground Truth	1.21	0.00	0.00
	SimpleBaseline	10.14	11.60	8.13
	+ SEAL-Pose (margin)	**9.30**	**8.39**	**7.19**
	+ SEAL-Pose (NCE)	10.30	8.84	7.35
H3WB	Ground Truth	4.42	0.00	0.00
	SimpleBaseline	**6.60**	6.56	6.22
	+ SEAL-Pose (margin)	6.78	**6.24**	**5.86**
	+ SEAL-Pose (NCE)	6.85	6.32	5.98

On the H3WB dataset, SEAL-Pose showed mixed results, with better LLE and BSLE but higher LSE compared to the baseline. This is likely due to the dataset's noisy labeling, which is apparent from the high LSE of the ground truth poses. It is probable that SEAL-Pose would struggle to improve limb symmetry on H3WB with noisy and asymmetric labeling, as the loss-net may not be able to learn and provide additional guidance about it. Indeed, the trained task-nets also exhibited similar levels of LSE with the ground truth poses.

Overall, the improved structural consistency metrics highlight that loss-net's ability to capture structures in human poses helps the task-net to predict more anatomically consistent and plausible 3D human poses.

6.3. Analysis of Structured Energy Network

Gradient-Based Inference Analysis We conduct gradient-based inference (GBI) on the output of the task-net using the trained loss-net to verify its ability to capture plausible human pose structures. GBI iteratively refines the predicted poses by following the gradient signals from the loss-net, which are expected to lower the assigned energy. As shown in Figure 3, P-MPJPE, as well as the structural metrics LSE, LLE, and BSLE, all steadily decrease over dozens of iterations. Since these metrics directly reflect structural plausibility, the consistent reduction indicates that the loss-net effectively caputres human pose structure and provides meaningful gradient signals. The effect is more pronounced on the challenging 3DHP dataset but the same trend is also observed on H36M, as shown in Figure 5 in Appendix A, confirming the consistency of the results.

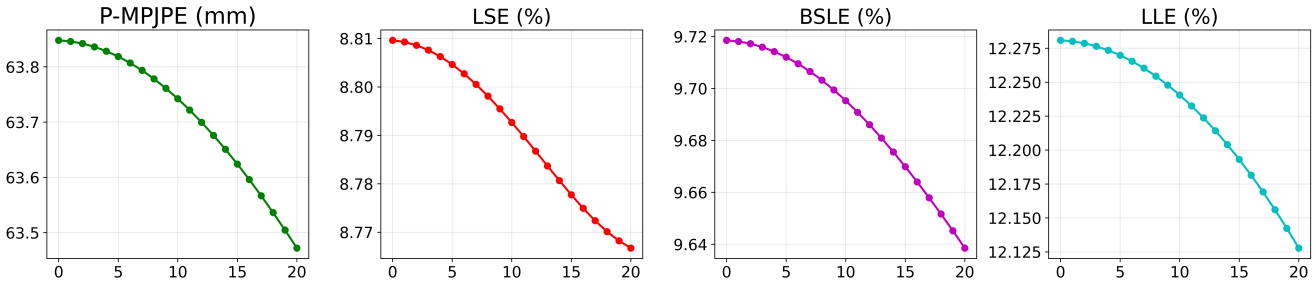

Figure 3. **Gradient-Based Inference results on MPI-INF-3DHP**. P-MPJPE, LSE, LLE, and BSLE all decrease steadily over iterations, indicating that the loss-net effectively captures structural plausibility and provides meaningful corrective feedback to the task-net.

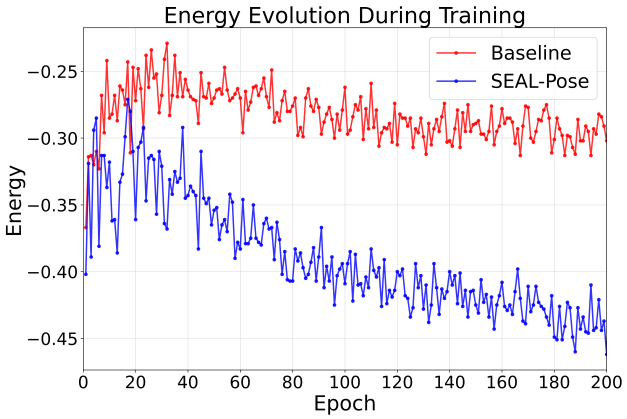

Figure 4. **Training Dynamics of Energy**. Energy evolution during training shows SEAL-Pose on H3WB, effectively lowers the energy level of task-net predictions.

Energy Evolution during Training To evaluate the efficacy of SEAL-Pose in exploiting the learning signals from the loss-net, we observed the energy levels of the task-net predictions at each training checkpoint, logged at every epoch. We calculated the energy of the task-net predictions by averaging the energy values of the predicted poses across the entire test set of H3WB and compared the energy levels between the baseline model and SEAL-Pose. During the training process, the task-net in SEAL-Pose demonstrated a more significant reduction in energy levels compared to the baseline, resulting in considerably lower energy values at the end of the training, as shown in Figure 4. The decrease in energy suggests that SEAL-Pose effectively utilizes the structured energy network to teach the task-net towards more plausible predictions by guiding the task-net to lower the energy of the predicted poses. This observation supports the hypothesis that the loss-net can train the task-net as we intended and strongly suggests that it guides the task-net effectively, given our prior demonstration that structured energy networks capture plausible pose structures.

7. Conclusion

In this work, we introduced SEAL-Pose, a novel adaptation of the SEAL framework to deterministic models, particularly in 3D human pose estimation. Our approach employs a structured energy network as a trainable loss function, effectively capturing joint dependencies and improving the plausibility of predicted poses without any explicit structural priors. In addition, we suggest new metrics: limb symmetry error (LSE) and body segment length error (BSLE) to quantitatively evaluate the structural consistency of the generated poses. Our experimental results showed the effectiveness of SEAL-Pose over baselines, achieving substantial reductions in per-joint errors. SEAL-Pose also demonstrated better structural consistency, as evidenced by lower LSE and BSLE values, which underscore the efficacy of SEAL-Pose in capturing complex structures among body joints. This work highlights the potential of structured energy networks for enhancing tasks involving complex output dependencies. Our findings suggest that SEAL-Pose can be extended to broader applications in the future, providing a promising direction to model the complex dependencies in high-dimensional output spaces, thereby improving the performance and structural consistency in various domains.

8. Limitations

Although SEAL-Pose demonstrates significant improvements in 3D human pose estimation, there are still room for advancement. One key challenge lies in the broad hyperparameter search space, which includes weight for the energy loss term, learning rates for the task-net and the loss-net, relative batch size of the task-net and loss-net and such. This extensive search space can make the training optimization process less straightforward and computationally intensive. Therefore, analyzing and identifying efficient strategies for robust and stable training could enhance the practicality of the method. Additionally, better loss-net architectures to provide learning signals for the task-net could further improve the joint optimization process of SEAL-Pose.

References

[1] David Belanger and Andrew McCallum. Structured prediction energy networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 983–992. JMLR.org, 2016. 2, 4

[2] Mosam Dabhi, László A. Jeni, and Simon Lucey. 3d-lfm: Lifting foundation model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10466–10475, 2023. 6

[3] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6821–6828. AAAI Press, 2018. 1, 2, 3

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 262–270, 2015. 4

[5] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 5

[8] Jun-Hee Kim and Seong-Whan Lee. Toward approaches to scalability in 3d human pose estimation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 3

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[10] Jay Yoon Lee, Sanket Vaibhav Mehta, Michael L. Wick, Jean-Baptiste Tristan, and Jaime G. Carbonell. Gradient-based inference for networks with output constraints. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 4147–4154. AAAI Press, 2019. 4

[11] Jay-Yoon Lee, Dhruvesh Patel, Purujit Goyal, Wenlong Zhao, Zhiyang Xu, and Andrew McCallum. Structured energy network as a loss. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4

[12] Yang Liu, Changzhen Qiu, and Zhiyong Zhang. Deep learning for 3d human pose estimation and mesh recovery: A survey. *Neurocomputing*, page 128049, 2024. 1, 2, 5

[13] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3698–3707, Brussels, Belgium, 2018. Association for Computational Linguistics. 4

[14] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. 5

[15] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 5

[16] A. Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 4

[17] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[18] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[19] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[20] Cédric Rommel, Victor Letzelter, Nermin Samet, Renaud Marlet, Matthieu Cord, Patrick Pérez, and Eduardo Valle. Manipose: Manifold-constrained multi-hypothesis 3d human pose estimation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024. 2

[21] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 5

[22] Lele Wu, Zhenbo Yu, Yijiang Liu, and Qingshan Liu. Limb pose aware networks for monocular 3d pose estimation. *IEEE Transactions on Image Processing*, 31:906–917, 2022. 1, 2, 3

[23] Yuanlu Xu, Wenguan Wang, Tengyu Liu, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6327–6344, 2022. 1, 2, 3

[24] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, 2022. 5

[25] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3425–3435, 2019. 5

[26] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1), 2023. 2, 5

[27] Xiangtao Zheng, Xiumei Chen, and Xiaoqiang Lu. A joint relationship aware neural network for single-image 3d human pose estimation. *IEEE Transactions on Image Processing*, 29: 4747–4758, 2020. 1, 2, 3

[28] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177, 2023. 2, 5

A. Gradient-Based Inference

The gradient-based inference results on the H36M dataset are shown in Figure 5, where all metrics steadily decrease over iterations, consistent with the trends observed for 3DHP in Figure 3.

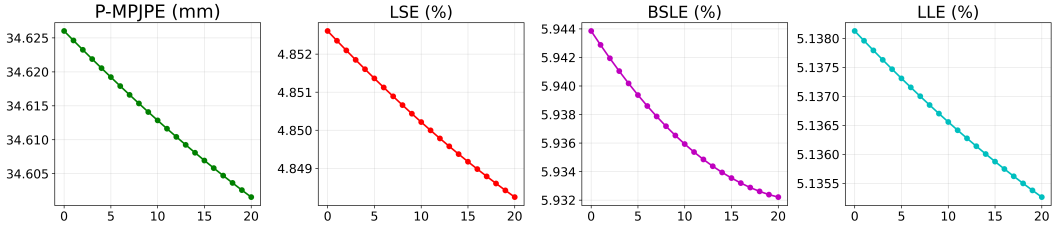

Figure 5. **Gradient-Based Inference results on H36M**. P-MPJPE, LSE, LLE, and BSLE all decrease steadily over iterations, indicating that the loss-net effectively captures structural plausibility and provides meaningful corrective feedback to the task-net.