FusionFormer: A Concise Unified Feature Fusion Transformer for 3D Pose Estimation

Yanlu Cai[1], **Weizhong Zhang**[1,2,*], **Yuan Wu**[1], **Cheng Jin**[1,2,*]

[1]Fudan University, Shanghai, China
[2]Innovation Center of Calligraphy and Painting Creation Technology, MCT, China
{ylcai20, weizhongzhang, wuyuan, jc}@fudan.edu.cn

Abstract

Depth uncertainty is a core challenge in 3D human pose estimation, especially when the camera parameters are unknown. Previous methods try to reduce the impact of depth uncertainty by multi-view and/or multi-frame feature fusion to utilize more spatial and temporal information. However, they generally lead to marginal improvements and their performance still cannot match the camera-parameter-required methods. The reason is that their handcrafted fusion schemes cannot fuse the features flexibly, e.g., the multi-view and/or multi-frame features are fused separately. Moreover, the diverse and complicated fusion schemes make the principle for developing effective fusion schemes unclear and also raises an open problem that whether there exist more simple and elegant fusion schemes. To address these issues, this paper proposes an extremely concise unified feature fusion transformer (FusionFormer) with minimized handcrafted design for 3D pose estimation. FusionFormer fuses both the multi-view and multi-frame features in a unified fusion scheme, in which all the features are accessible to each other and thus can be fused flexibly. Experimental results on several mainstream datasets demonstrate that FusionFormer achieves state-of-the-art performance. To our best knowledge, this is the first camera-parameter-free method to outperform the existing camera-parameter-required methods, revealing the tremendous potential of camera-parameter-free models. These impressive experimental results together with our concise feature fusion scheme resolve the above open problem. Another appealing feature of FusionFormer we observe is that benefiting from its effective fusion scheme, we can achieve impressive performance with smaller model size and less FLOPs.

Introduction

3D human pose estimation (Wang et al. 2021) is a fundamental task in computer vision, which aims to estimate 3D locations of the keypoints on the human body from images or videos. Although great efforts have been made in the last decade, it remains challenging due to the depth uncertainty. Recent approaches (Zheng et al. 2021; Liu et al. 2021; Iskakov et al. 2019; Zhang et al. 2021b) try to reduce depth uncertainty by leveraging the clues contained in the features from multiple views and frames. Promising results have been reported in the literature. To be precise, the multi-frame methods, such as PoseFormer (Zheng et al. 2021) and MHFormer (Li et al. 2022), take advantage of Transformer's strong capability in long-range relationship modelling to extract robust features and reduce the impact of the inaccuracies in 2D pose estimation. The multi-view methods (Iskakov et al. 2019; Ma et al. 2021; He et al. 2020) fuse the features of the images from multiple views via geometric constraints, which can be derived from the camera parameters. There also exist some camera-parameter-free methods (Gordon et al. 2022; Shuai, Wu, and Liu 2022). They leverage Transformers to infer the camera parameters explicitly or implicitly. Moreover, these camera-parameter-free methods do not require image or voxel features for geometric alignment, which leads to significant computational cost savings and enables these multi-view methods to jointly learn from multiple frames instead of single frame as the camera-parameter-required methods.

In this paper, we focus on the camera-parameter-free approaches as they are more practical in real applications. However, we notice that the performance of these approaches (Ma et al. 2022; Gordon et al. 2022; Shuai, Wu, and Liu 2022) can not match the camera-parameter-required methods although they are able to learn from multiple frames. We argue that this results from their complex handcrafted feature fusion schemes, which cannot fuse the features flexibly. For instance, grouping keypoint feature fusion scheme first groups the keypoints within the same limb and then fuses the features separately for each group, which hinders the cooperative interactions between limbs. The pairwise feature fusion scheme may suffer from a lack of global perspective and the reference to other views. As a result, common features may be repeatedly extracted while cetrain informative features may be overlooked. Spatio-temporal separated feature fusion scheme restricts the features from communicating along either in the spatial or temporal dimension. Moreover, we find that the above handcrafted fusion schemes have diverse and complex structures, which make the principle for developing effective fusion scheme unclear. They also raise an open problem that *whether there exist more simple, elegant yet effective fusion schemes*.

To address the above issues, this paper proposes a concise unified Fusion Transformer (FusionFormer) with minimized handcrafted design for 3D pose estimation. FusionFormer

*Corresponding authors

first encodes 2D pose estimation results into pose features, and then leverages transformer encoder to jointly fuse multi-view and multi-frame features into the global feature. Thus, all the features are accessible to each other and thus can be fused flexibly. Subsequently, Transformer decoder is adopted to estimate 3D human pose under each view by integrating the global features with the view-specific features individually. The extensive experimental results demonstrate that our method outperforms the state-of-the-art methods with a large margin. To the best of our knowledge, FusionFormer[1] is the first camera-parameter-free method to surpass existing camera-parameter-required methods benefiting from our effective fusion scheme, revealing the tremendous potential of camera-parameter-free models. Furthermore, the experimental results demonstrate that when using 2D ground truth as input, our method achieves extremely high accuracy with MPJPE error less than 10mm. This implies that with more accurate 2D pose estimation techniques our method can be further reinforced in the future. The impressive experimental results together with our concise feature fusion scheme resolve the above open problem. Our main contributions are summarized as follows:

1. We propose a concise unified feature fusion transformer (FusionFormer) for 3D pose estimation, which enables the features to be fused from different frames and views.
2. Experiments on several datasets demonstrate that FusionFormer outperforms all the state-of-the-art methods, with an accurracy improvement more than 23%. This demonstrates that the impact of depth uncertainty is effectively reduced by FusionFormer, revealing the tremendous potential of camera-parameter-free methods.
3. To the best of our knowledge, FusionFormer is the first camera-parameter-free method that surpasses existing camera-parameter-required methods benefiting from our effective fusion scheme.
4. The success of FusionFormer verifies the existence of concise yet effective 3D pose estimation approaches, which can inspire the researchers in developing more advanced approaches in the future.

Related Work

Monocular 3D Pose Estimation

Recovering 3D pose information from a single view is an ill-posed problem. Therefore, single-view methods usually introduce the prior knowledge of the human body to constrain the location of keypoints, reducing the depth uncertainty. Common constraints include reprojection, e.g., Occlusion-aware Network (OA-Net) (Cheng et al. 2019), relationship between adjacent keypoints or within the same limb, e.g., SemGCN (Zhao et al. 2019), HDFormer (Chen et al. 2023) and SRNet (Zeng et al. 2020), human body symmetry, e.g., PoseGrammar (Fang et al. 2018), bone length invariance, e.g., MotioNet (Shi et al. 2020), temporal consistency, e.g., VideoPose3D (Pavllo et al. 2019), and temporal motion constraints, e.g., UGCN and GASTNet (Liu et al. 2021).

[1]Code and Supplementary materials are available at: https://github.com/DoUntilFalse/FusionFormer

Before the Transformer was introduced into 3D Human Pose Estimation, monocular methods used various handcrafted schemes to encode prior knowledge. However, in these methods, the keypoint features can only be fused along a pre-designed graph, which limits the exploration of relationships between two non-adjacent keypoints in the graph.

PoseFormer (Zheng et al. 2021) proposes the first Transformer-based method, and reveals the powerful potential of the Transformer in 3D pose estimation. Inspired by MDN (Li and Lee 2019) and other Multi-Hypothsis methods (Jahangiri and Yuille 2017; Liu et al. 2023), MHFormer (Li et al. 2022) proposes a multi-hypothesis transformer, which first generates multiple possible 3D poses and then fuses the final 3D pose to reduce depth uncertainty.

Multi-view 3D Pose Estimation

Camera-parameter-required Approaches. Multi-view methods with camera parameters model the position relationships between cameras to construct epipolar geometry constraints, reducing depth uncertainty and the impact of occlusion. Learnable Triangulation of Human Pose (LToHP) (Iskakov et al. 2019), CanonFusion (Remelli et al. 2020) and other studies (Li et al. 2019; Günel et al. 2019) propose various triangulation methods to determine the 3D keypoint positions. DeepFuse (Huang et al. 2020) and CrossFusion (Qiu et al. 2019) reproject the 2D feature into 3D voxel space. AdaFuse (Zhang et al. 2021b) and Epipolar Transformer (He et al. 2020) leverage epipolar line to fuse multi-view features. TransFusion (Ma et al. 2021), MvP (Zhang et al. 2021a) and MTF-Transformer+ (MTF+) (Shuai, Wu, and Liu 2022) feed camera parameters into their position encoding to guide the model in modeling the relationships between views.

Camera-parameter-free Approaches. Multi-PPT (Ma et al. 2022) employs Transformer to extract human features from images in multiple views, and then feeds them into a shared Transformer encoder for feature fusion to obtain 3D human poses. Recently, some methods attempt to utilize both multi-frame and multi-view feature fusion to boost the performance. FLEX (Gordon et al. 2022) designs a viewpoint-independent skeleton representation that uses skeleton lengths and angles to represent the human body and leverage the prior knowledge of invariant skeleton lengths. Notice that hierarchical keypoints representation leads to error accumulation, resulting in large errors at the end-point keypoints, such as wrist and ankle. MTF-Transformer (MTF) (Shuai, Wu, and Liu 2022) fuses keypoint features, multi-view features and multi-frame features separately to obtain more spatial and temporal information.

Method

In this section, we present our unified feature fusion transformer, dubbed FusionFormer, which jointly fuses features from multiple frames and views to estimate 3D human poses accurately. As shown in Figure 1, FusionFormer consists of four modules, i.e., 2D pose estimator, pose feature extractor, unified feature fusion scheme, and 3D pose regression head. We would like to point out that our feature fusion scheme

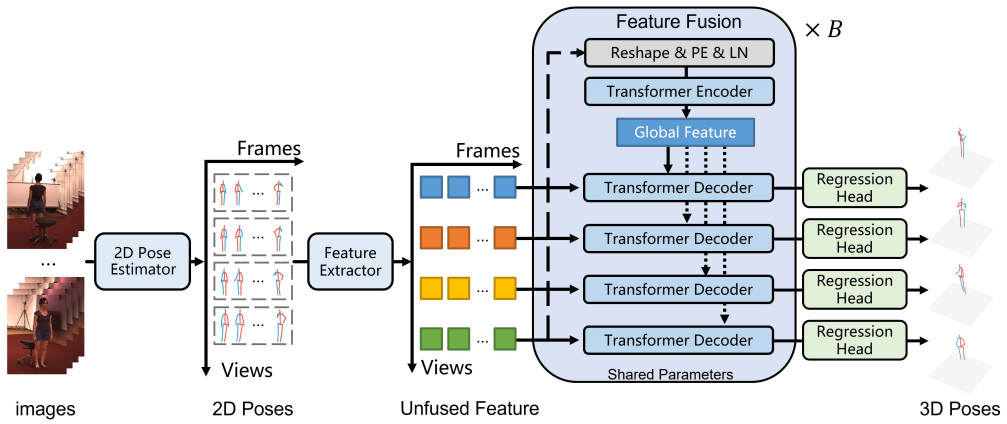

Figure 1: The architecture of FusionFormer. FusionFormer decompose the 3D Human pose estimation into four stage: 2D Pose estimation, Feature Extraction, Feature Fusion and 3D Pose Regression Head. FusionFormer unifies spatial and temporal feature fusion to further exploit the powerful modeling capabilities of Transformer.

can be integrated with existing techniques for other three modules flexibly and therefore we only provide a brief introduction to the formal expressions of these three modules.

2D Pose Estimator and Feature Extractor

A general 2D pose estimator, which estimates the poses in the images $\mathcal{I}$ obtained from V views with T frames in each view, can be formulated as the following function

$$F_{2D} : \mathbb{R}^{T\times V\times H\times W\times 3} \rightarrow \mathbb{R}^{T\times V\times J\times 2},$$

where J are the number of keypoints for one person, H and W are height and width of each image. We denote the estimation result as $\mathcal{P}_{2D}$, that is

$$\mathcal{P}_{2D} = F_{2D}(\mathcal{I}) \in \mathbb{R}^{T\times V\times J\times 2}.$$

Feature extractor takes the 2D poses $\mathcal{P}_{2D}$ as input and maps it into a high-dimensional space to obtain the feature $\mathcal{F}_{embed} \in \mathbb{R}^{T\times V\times J\times C_J}$, that is

$$\mathcal{F}_{embed} = Embed(\mathcal{P}),$$

where C_J is the number of channels of each keypoint.

Subsequently, the feature extractor employs several layers to extract the relationships between keypoints, resulting in the pose feature $\mathcal{F}^{(0)}_{pose} \in \mathbb{R}^{T\times V\times C_P}$ with C_P being the number of channels of each pose, which aggregates the features of all J keypoints. That is,

$$\mathcal{F}^{(0)}_{pose} = E_{pose}(\mathcal{F}_{embed}).$$

Unified Feature Fusion Scheme

We argue that the features from multiple views and frames should be accessible from each other in the feature fusion process to reduce the impact of depth uncertainty. Therefore, we propose a unified feature fusion scheme, which is composed of several Encoder-Decoder blocks. The encoder is used to fuse all the features of VT images to obtain a global feature $\mathcal{F}_{global}$. The decoder integrates the global features with the view-specific features individually to provide more informative features for 3D human pose under each view. The impact of depth uncertainty can be reduced as our fused features have the global perspective. The details of our encoders and decoders are presented below.

Encoder. Before being fed into the encoder, the feature $\mathcal{F}^{(0)}_{pose}$ needs to be reshaped and position encoding needs to be added. To perform feature communication and fusion across both spatial and temporal dimensions with Transformer, we treat $\mathcal{F}^{(0)}_{pose}$ as $V * T$ tokens. Common position encodings include cosine position encoding, learnable position encoding and MLP-based position encoding. We note that cosine position encoding (Vaswani et al. 2017) assumes that there is correlation among tokens that decreases with distance, which may not hold true in multi-view and multi-frame feature fusion. Moreover, MLP-based position encoding is mainly suitable for scenarios where the number of tokens changes dynamically. When the number of tokens is fixed, it will degrade to a learnable position encoding. Therefore, inspired by ViT (Dosovitskiy et al. 2021), we adopt a simple learnable position encoding. After position encoding were added, Layer Normalization are applied. The whole process can be formulated as

$$\mathcal{F}^{(0)}_{enc} = LN(f(\mathcal{F}^{(0)}_{pose}) + PE_{enc}),$$

where f reshapes the feature into $V * T$ tokens, $PE_{enc} \in \mathbb{R}^{(V*T)\times C_P}$ is the learnable position encoding and LN is Layer Normalization.

$\mathcal{F}^{(0)}_{enc}$ is then fed into L layers of Transformer Encoder to obtain the global feature $\mathcal{F}^{(0)}_{global}$, i.e.,

$$\mathcal{F}^{(0)}_{global} = Encoder(\mathcal{F}^{(0)}_{enc}).$$

To clarify, when referring to the Transformer Encoder or Decoder in this paper, we adopt the vanilla Transformer (Vaswani et al. 2017), and we will not elaborate on its specific structure.

As we know, Transformer Encoder is composed of two key modules: a self-attention module and an MLP module. The self-attention module is employed for feature fusion among tokens to obtain the global feature, whereas the MLP module serves for inter-channel feature communication. One advantage of flattening features across V views and T frames into a sequence is that direct communication can be performed between any feature pair from the $V*T$ features. Consequently, the attention matrix is sized of $VT \times VT$ instead of two matrices sized of $V \times V$ and $T \times T$ as previous approaches. We would like to point out that as 1) V in the typical datasets (Ionescu et al. 2013) is no larger than 4; 2) our Frames $T = 27$ and Blocks $B = 2$ are significantly smaller than the baselines (e.g., $T = 81$ and $B = 8$ in PoseFormer (Zheng et al. 2021), $T = 351$ and $B = 3$ in MHFormer (Li et al. 2022)), our computation and memory cost is comparable with the baselines.

Moreover, this design endows each unit pair with an independent attention weight, leading to more effective and flexible feature fusion compared to the separated fusion methods that rely on shared weights for every view in multi-frame feature fusion or every frame in multi-view feature fusion.

In contrast to pairwise multi-view feature fusion methods, where attention weights are normalized separately, a uniform normalization for our attention weights is performed in the self-attention module. This allows for different sums of attention weights across the views, therefore, we allow different views to have different importance weights in the fused feature. In pairwise methods, the sum of attention weights for every pair of views remains constant due to the separate normalization, regardless of their importance.

In practice, the relationships between the V views often differ greatly due to factors such as occlusion (Ghafoor and Mahmood 2022), camera position, body orientation, and other factors, making it difficult to manually summarize them. This also explains why previous handcrafted structures struggle to achieve optimal results.

Decoder. As our goal is to predict the 3D pose in V views, we try to maintain the diversity between features from different views, so we partition the features $\mathcal{F}_{pose}^{(0)}$ along the views and then feed them separately into the decoder. Therefore, multi-view feature fusion is only performed in the encoder. To be precise, the features are first partitioned and normalized together with the positional encoding as follows:

$$\{\mathcal{F}_{dec}^{(0)}\}_v = LN(\{\mathcal{F}_{pose}^{(0)}\}_v + PE_{dec}).$$

Then global features are integrated with the above view-specific features $\{\mathcal{F}_{dec}^{(0)}\}_v$ individually in our decoder with L layers to provide more informative features for 3D human pose under each view. That is

$$\{\mathcal{F}_{fused}^{(0)}\}_v = Decoder(\mathcal{F}_{global}^{(0)}, \{\mathcal{F}_{dec}^{(0)}\}_v).$$

After these operations, the features for each view are concatenated into $\mathcal{F}_{fused}^{(0)} \in \mathbb{R}^{V \times T}$.

Encoder-Decoder Block. The aforementioned Encoder-Decoder is treated as a single block. As shown in Figure 1, FusionFormer contains B blocks with shared parameters, where the output $\mathcal{F}_{fused}^{(b)}$ of the b-th block serves as the input of the subsequent block. Finally, the output of this stage is obtained as $\mathcal{F}_{fused}^{(B)}$.

3D Pose Regression Head and Loss Function

To maximize the capabilities of our feature fusion network, we employ a simple 3D pose regression head to extract the 3D poses of the center frame (the $\frac{T+1}{2}$-th frame) from each view separately following MTF-Transformer. We adopt a Conv1d layer to perform the weighted summation of all frames, aggregating information from all frames to obtain features $\mathcal{F}_{agg} \in \mathbb{R}^{V \times C_P}$ that are used for 3D pose regression, i.e.,

$$\mathcal{F}_{agg} = Conv1d(\mathcal{F}_{pose}^{(B)}).$$

Afterward, $\mathcal{F}_{agg}$ is fed into two linear layers to obtain the 3D poses of the center frame for each view, denoted as $\mathcal{P}_{3D} \in \mathbb{R}^{V \times J \times 3}$, i.e.,

$$\mathcal{P}_{3D} = Linear(Relu(Linear(\mathcal{F}_{agg}))).$$

We adopt Mean Per Joint Position Error (Wang et al. 2021) as our loss function and denote it as MPJPE. MPJPE first aligns the root (central hip) of predicted 3D pose and the ground truth, and then calculates the averaged Euclidean Distance between each joints. We adopt the averaged MPJPE over V views as the final loss function, i.e.,

$$\mathcal{L} = \frac{1}{V * J} \sum_{v=1}^{V} \sum_{j=1}^{J} \left\| p_{v,j} - p_{v,j}^{gt} \right\|,$$

where $p_{v,j} \in \mathcal{P}_{3D}$ represents the predicted 3d poses after alignment, and $p_{v,j}^{gt}$ represents the ground truth.

Remark. *We adopt PoseFormer as the feature extractor in the main experiments. Note that FusionFormer achieves impressive results even with a 2-layer FCN (Table 7). It verifies that FusionFormer is not a simple extension of PoseFormer.*

Experiments

Experiment Setting

Dataset. Human3.6M is the most widely used 3D human pose estimation dataset, containing over 3 million frames of images synchronized captured from four cameras. TotalCapture dataset utilizes 8 completely synchronized cameras to collect 4 types of actions (rom, acting, walking, and freestyle) from 5 subjects (S1, S2, and S3 as Seen subjects and S4 and S5 as Unseen subjects). HumanEva ($\approx$ 50K frames) and MPI-INF-3DHP ($\approx$ 500K frames) are two much smaller datasets. HumanEva contains 3 calibrated rgb video sequences from 4 subjects performing 6 common actions. MPI-INF-3DHP consists of both constrained indoor and complex outdoor scenes captured from 14 cameras.

Evaluation Metrics. Mean Per Joint Position Error (MPJPE) and Procrustes-aligned MPJPE (P-MPJPE) (Wang et al. 2021) are used as the evaluation metrics.

Method	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit.	Smoke	Wait	Walk	Avg.
Monocular methods													
MDN* (T = 1)	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	54.8	50.6	43.4	52.7
SRNet* (T = 243)	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	46.1	42.6	31.5	44.8
UGCN* (T = 96)	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	46.3	42.2	31.7	44.5
PoseFormer* (T = 81)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	45.5	43.3	31.8	44.3
MHFormer* (T = 351)	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	43.7	41.1	29.8	43.0
OA-Net* (T = 128)	38.3	41.3	46.1	40.1	41.6	51.9	41.8	40.9	51.5	42.2	44.6	33.7	42.9
Multi-view methods with camera parameters													
CanonFusion (T = 1)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	31.2	29.9	33.7	30.2
Epipolar (T = 1)	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	28.2	26.4	28.3	26.9
CrossFusion (T = 1)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	31.0	25.6	28.0	26.2
TransFusion (T = 1)	24.4	26.4	23.4	21.1	25.2	23.2	24.7	33.8	29.8	26.8	24.2	26.1	25.8
LToHP (T = 1)	19.9	20.0	18.9	18.5	20.5	19.4	18.4	22.1	22.5	21.2	20.8	22.1	20.8
AdaFuse (T = 1)	17.8	19.5	17.6	20.7	19.3	16.8	18.9	20.2	25.7	19.2	20.5	20.5	19.5
MvP (T = 1)	-	-	-	-	-	-	-	-	-	-	-	-	18.6
MTF+* (T = 27)	23.4	25.2	23.1	24.4	27.4	28.5	22.8	25.2	28.7	25.9	23.6	22.6	25.8
Multi-view methods without camera parameters													
Multi-PPT (T = 1)	21.8	26.5	21.0	22.4	23.7	23.1	23.2	27.9	30.7	26.7	23.3	25.3	24.4
FLEX* (T = 27)	-	-	-	-	-	-	-	-	-	-	-	-	31.7
MTF* (T = 27)	23.1	25.4	24.7	24.5	27.9	28.3	23.9	24.6	30.7	25.8	24.2	22.8	26.2
MTF† (T = 27)	17.6	21.3	15.0	18.6	17.6	23.9	16.5	16.9	17.5	18.5	17.0	15.4	18.7
Ours*(T=27)	22.2	25.3	22.9	23.6	26.0	27.0	22.4	23.9	30.4	25.6	22.8	22.3	25.4
Ours† (T = 27)	**15.7**	**15.6**	**13.0**	**15.9**	**13.9**	**15.6**	**14.9**	**15.5**	**15.5**	**14.3**	**15.2**	**14.6**	**15.1**
FLEX‡ (T = 27)	-	-	-	-	-	-	-	-	-	-	-	-	22.9
MTF‡ (T = 27)	15.5	17.1	13.7	15.5	14.0	16.2	15.8	16.5	15.8	14.5	14.5	14.3	15.3
Ours‡ (T = 27)	7.84	8.04	7.39	8.33	7.13	9.02	8.00	8.19	7.57	7.37	7.83	7.26	7.90

Table 1: Results on Human3.6M. MPJPE is adopted as the evaluation metric. We adopt CPN (*) and ViTPose (†) as the 2D pose estimator for fair comparison. T is the number of frames. We adopt 2DGT (‡) as input to explore the theoretical upper bound of the model. Due to space constraints, we only report detailed results for partial actions.

2D Pose Estimator. We employ two off-the-shelf 2D pose estimators CPN (Chen et al. 2018) and ViTPose (Xu et al. 2022) on Human3.6M for fair comparison with 2D-to-3D methods and images-to-3D methods, respectively. Detailed considerations are given in Main Results section. Following MTF-Transformer, we employ ResNet101 (He et al. 2015) as the 2D pose estimator on TotalCapture dataset.

Feature Extractor. We choose PoseFormer as the pose feature extractor, denoted as E_P, for the main experiment. To show the flexibility of FusionFormer in integrating with feature extractors, we propose 2 baselines, i.e. E_{FC} and E_T. E_{FC} is a 3-layer FCN. E_T is a 2-layer FCN followed by 2 vanilla Transformer layers with learnable position encoding.

More detailed configuration, e.g., dataset partitioning and learning rate, is postponed to the supplementary materials.

Main Results

We report the general comparison results on Human3.6M and TotalCapture and the generalization ability evaluation results on two small datasets HumanEva and MPI-INF-3DHP. Moreover, to show the superiority of FusionFormer further, we give more comparison results with the previous state-of-the-art method MTF-Transformer on extra aspects, including scalability, computational efficiency and visual analysis. Finally, we reveal the reasons behind the superiority of FusionFormer by designing a camera extrinsic parameter regression task as well as visualizing attention maps.

Human3.6M. Note that 2D-to-3D methods use CPN as the 2D pose estimator, whereas image-to-3D methods usually adopt advanced structures, e.g., Transformer, to achieve higher accuracy. For fair comparison, we report the results of FusionFormer with CPN or ViTPose as the pose estimator. We replace 2D pose estimation with 2D ground truth (2DGT) to explore the theoretical upper bound of our method. All the results are given in Table 1.

We can observe that FusionFormer outperforms both the 2D-to-3D and image-to-3D methods with large margins, i.e., 25.3mm *v.s.* 26.2mm (MTF-Transformer) and 31.7mm (FLEX) with CPN, 15.1mm *v.s.* 18.6mm (MvP) and 24.4mm (PPT) with Transformer-based structures.

Notably, Table 1 demonstrates that FusionFormer with CPN outperforms some camera-parameter-required methods (e.g., MTF-Transformer+ and Cross-view Fusion) and some methods with advanced Transformer based structures (e.g., Epipolar Transformer and TransFusion). When ViTPose is adopted, it can beat all the methods consistently.

Moreover, Table 1 shows that when 2D pose estimation results are replaced with 2DGT, the performance of FusionFormer is significantly boosted. Precisely, it achieves an MPJPE of 7.91mm, far lower than other 2D-to-3D methods, e.g., MTF-Transformer (15.3mm) and FLEX (22.9mm). This implies that the accuracy of the 2D pose estimator has become the main bottleneck in accurate 3D pose estimation. It is also worth noting that the MPJPE of FusionFormer with ViTPose is even lower than the theoretical upper bound

Method	Seen Cameras(1,3,5,7)							Unseen Cameras(2,4,6,8)							Mean
	Seen Subjects			Unseen Subjects			Mean	Seen Subjects			Unseen Subjects			Mean	
	W2	FS3	A3	W2	FS3	A3		W2	FS3	A3	W2	FS3	A3		
CrossFusion(*)†	19.0	28.0	21.0	32.0	54.0	33.0	29.0	-	-	-	-	-	-	-	-
CanonFusion(*)†	10.6	30.4	16.3	27.0	65.0	34.2	27.5	22.4	47.1	27.8	39.1	75.7	43.1	38.2	32.9
MTF+(Res101)†	10.7	26.5	16.7	27.4	49.4	34.1	25.1	**13.9**	**29.2**	**18.1**	29.2	49.5	35.6	**27.0**	26.1
FLEX(Res101)	33.2	81.0	34.2	38.3	124	59.5	49.4	109	152	105	114	176	123	125	87.4
MTF(Res101)	9.30	26.5	14.5	26.7	53.1	33.8	24.7	23.7	40.3	27.4	37.0	61.8	42.9	36.6	30.7
Ours(Res101)	**5.50**	**15.0**	**5.68**	**18.1**	**37.6**	**20.6**	**15.0**	22.1	35.4	23.4	**23.2**	**42.6**	28.4	28.3	**21.7**

Table 2: Comparison results on TotalCapture. MPJPE is adopted as the evaluation metric. We adopt Res101 (He et al. 2015) as the 2D pose estimator for fair comparison. † marks methods that require camera parameters as input.

Datasets	HumanEva	MPI-INF-3DHP
Poseformer(T=27)	35.9	38.5
MTF-Transformer(T=3)	22.8	14.6
Ours(T=3)	15.4	5.4

Table 3: Generalization experiment. 2DGT are used as input.

Methods	Axis	Angle	Trans.
PoseFormer	20°	11°	1.1m
MTF-Transformer	6°	5°	0.3m
Ours	$\mathbf{4}^\circ$	$\mathbf{2}^\circ$	**0.2m**

Table 4: Average error of Camera Extrinsic Parameter Regression.

of MTF-Transformer, i.e., 15.1mm *v.s.* 15.3mm, strongly demonstrating the superiority of our feature fusion scheme.

Finally, to verify that the superiority of FusionFormer comes from effective fuison scheme instead of 2D pose estimator ViTPose, we re-implement MTF-Transformer using ViTPose as 2D pose estimator. The results verify that FusionFormer still achieves lower error (15.1mm *v.s.* 18.7mm).

TotalCapture. We report the results across seen/unseen camera and seen/unseen subjects in Table 2. Due to the significant variations in freestyle videos, the freestyle action can be considered as an unseen action.

FusionFormer demonstrates better performance than all camera-parameter-free methods and outperforms the camera-parameter-required methods in the majority of scenarios, including the most challenging task with unseen cameras and unseen subjects. We notice that FusionFormer can not perform better than MTF-Transformer+, which is a camera-parameter-required method. It must be pointed that this is not a fair comparison, since with the given camera parameters as input, it can be expected that the impact of unseen cameras on camera-parameter-required methods is negligible. Nevertheless, the significant superiority of FusionFormer's over camera-parameter-required methods on the most challenging tasks with both unseen cameras and subjects demonstrates its impressive generalization ability.

Generalization across Datasets. We finetune a pretrained model on Human3.6M for 10 epochs on HumanEva or MPI-INF-3DHP, comparing our method's generalization capability with that of Poseformer and MTF-Transformer. All the models are trained with the same setting, which are given in the appendix. The results in Table 3 shows that FusionFormer outperforms the above two methods with a large margin, which indicates that our feature fusion scheme extracts robust feature representation in pretraining.

Scalability and Computational Efficiency. We report the results of MPJPE v.s. parameters/FLOPs in Figure 3 to show the superiority of FusionFormer in scalability and computational efficiency. We find that existing inflexible handcrafted fusion schemes lead to a low performance upper bound indicated with a dotted line in Figure 3(a), since some features can not been fused directly. To be precise, as shown in Figure 3(a), MTF-Transformer achieves the optimal performance at the model with 10M parameters, which is significantly larger than our model with only 1.89M parameters. Figure 3(b) shows that the small model size of FusionFormer finally leads to low computational complexity. Moreover, Figure 3 indicates that FusionFormer outperforms MTF-Transfromer with equal parameters/FLOPS.

Camera Extrinsic Parameter Regression. We conduct an experiment to evaluate the ability of models to predict camera position and orientation, which we refer to as Camera Extrinsic Parameter Regression. We freeze the parameters of each pre-trained model, extract the last layer features before the 3D pose regression head, and feed them into a 3-layer FCN to regress the camera extrinsic parameters, i.e., rotation axis, rotation angle, and translation. The results in Table 4 show that our method is more accurate than MTF-Transformer and PoseFormer. We speculate that the reason of why our method can accurately regress 3D human poses would stem from our ability to encode the relationships among cameras in feature fusion. All three methods perform poorly in predicting camera translation, which we attribute to the insensitivity of MPJPE to translation.

Visual Analysis. We visualize some results in Figure 2. It shows that FusionFormer performs well when self-occlusion and uncommon poses are present. In contrast, MTF-Transformer exhibits poor performance in such cases due to its handcrafted feature fusion scheme, which is consistent with our analysis in the introduction section.

Fusion Scheme Visualization. Figure 4 shows the attention matrix of FusionFormer to visualize our fusion scheme. In each block, the horizontal and vertical axes stand for 27 frames from two views. Thus, the 4 blocks in the diagonal

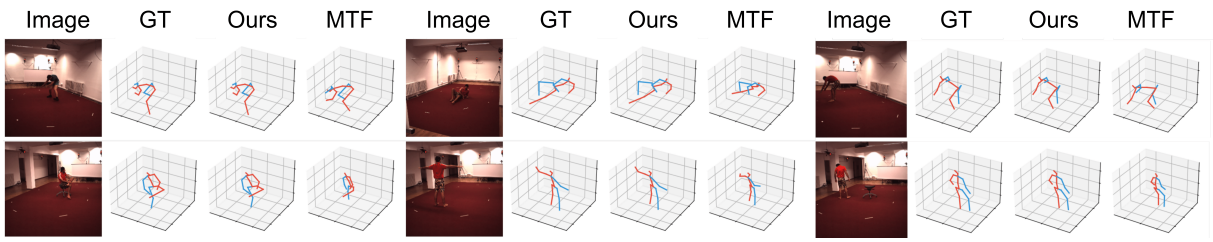

Figure 2: Results of MTF-Transformer and FusionFormer on Human 3.6M.

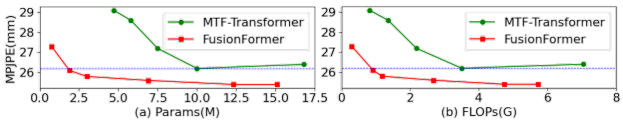

Figure 3: Comparison of MPJPE between our FusionFormer and MTF-Transformer over different model sizes.

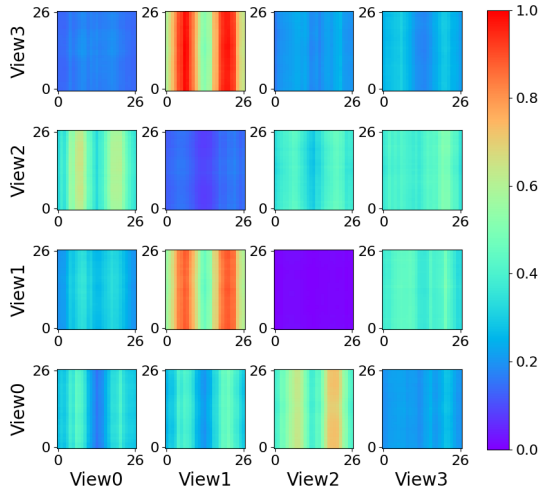

Figure 4: Visualization of attention map for feature fusion in FusionFormer decoder. The Attention map is divided into 4×4 blocks for better visualization. Each block contains the relationship between all frames (T=27) from a pair of views.

present the attention values of feature fusion among multiple frames in each view, while the diagonals in all 16 blocks are the values for fusion among multiple views. The large values out of the aforementioned 4 blocks and 16 diagonals stand for the feature pairs fused by FusionFormer, which come from different views and frames and are ignored by previous methods. Together with the superiority of FusionFormer, these large values imply that the fusion across the features from different views and frames is valuable and necessary.

Ablation Study

Frames and Views. The effects of frames and view numbers are shown in Table 5 and 6. They indicate that both temporal and multi-view information are helpful in reducing the impact of depth uncertainty. Additionally, more accurate 2D inputs require less temporal information.

Method	MPJPE		P-MPJPE	
	2DGT	CPN	2DGT	CPN
T = 81	8.77	25.7	4.64	20.8
T = 27	7.91	**25.3**	4.35	**20.5**
T = 9	**7.32**	26.2	**3.43**	21.3
T = 3	7.66	26.1	3.98	21.0
T = 1	8.87	27.3	4.93	21.7

Table 5: Ablation study for the number of frames T.

Method	MPJPE		P-MPJPE	
	2DGT	CPN	2DGT	CPN
V = 4	**7.91**	**25.3**	**4.35**	**20.5**
V = 2	13.1	30.8	8.26	24.9
V = 1	42.5	39.9	32.8	31.0

Table 6: Ablation study for the number of views V.

Method	MPJPE		P-MPJPE	
	2DGT	CPN	2DGT	CPN
E_{FC}	9.00	27.5	4.90	21.9
E_T	8.43	26.2	4.54	21.1
E_P	**7.91**	**25.3**	**4.35**	**20.5**

Table 7: Ablation study for Feature Extractor.

Feature Extractor. We evaluate the performance of FusionFormer with different feature extractors in Table 7. The results indicate that more complex feature extractors can indeed extract more discriminative features. However, even using just a few simple fully connected layers for feature extraction, FusionFormer achieves impressive performance, fully demonstrating the effectiveness of our scheme.

Conclusion

We propose a concise unified Feature Fusion Transformer for 3D pose estimation to reduce the impact of depth uncertainty, performing multi-view and multi-frame feature fusion in one step. Empirical results show that the superiority of our method with the accuracy improvement up to 23%.

Acknowledgments

This work was supported by National Natural Science Fund of China (62176064) and Shanghai Municipal Science and Technology Commission (22dz1204900).

References

Chen, H.; He, J.-Y.; Xiang, W.; Liu, W.; Cheng, Z.-Q.; Liu, H.; Luo, B.; Geng, Y.; and Xie, X. 2023. HDFormer: High-order Directed Transformer for 3D Human Pose Estimation. *arXiv preprint arXiv:2302.01825*.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; and Tan, R. T. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, 723–732.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fang, H.-S.; Xu, Y.; Wang, W.; Liu, X.; and Zhu, S.-C. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ghafoor, M.; and Mahmood, A. 2022. Quantification of occlusion handling capability of 3D human pose estimation framework. *IEEE Transactions on Multimedia*.

Gordon, B.; Raab, S.; Azov, G.; Giryes, R.; and Cohen-Or, D. 2022. FLEX: Extrinsic Parameters-free Multi-view 3D Human Motion Reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, 176–196. Springer.

Günel, S.; Rhodin, H.; Morales, D.; Campagnolo, J.; Ramdya, P.; and Fua, P. 2019. DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult Drosophila. *Elife*, 8: e48571.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.

He, Y.; Yan, R.; Fragkiadaki, K.; and Yu, S.-I. 2020. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 7779–7788.

Huang, F.; Zeng, A.; Liu, M.; Lai, Q.; and Xu, Q. 2020. Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 429–438.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36: 1325–1339.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7718–7727.

Jahangiri, E.; and Yuille, A. L. 2017. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 805–814.

Li, C.; and Lee, G. H. 2019. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9887–9895.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13147–13156.

Li, X.; Fan, Z.; Liu, Y.; Li, Y.; and Dai, Q. 2019. 3d pose detection of closely interactive humans using multi-view cameras. *Sensors*, 19: 2831.

Liu, H.; He, J.-Y.; Cheng, Z.-Q.; Xiang, W.; Yang, Q.; Chai, W.; Wang, G.; Bao, X.; Luo, B.; Geng, Y.; et al. 2023. PoSynDA: Multi-Hypothesis Pose Synthesis Domain Adaptation for Robust 3D Human Pose Estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5542–5551.

Liu, J.; Rojas, J.; Li, Y.; Liang, Z.; Guan, Y.; Xi, N.; and Zhu, H. 2021. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 3374–3380. IEEE.

Ma, H.; Chen, L.; Kong, D.; Wang, Z.; Liu, X.; Tang, H.; Yan, X.; Xie, Y.; Lin, S.-Y.; and Xie, X. 2021. TransFusion: Cross-view Fusion with Transformer for 3D Human Pose Estimation. In *British Machine Vision Conference*.

Ma, H.; Wang, Z.; Chen, Y.; Kong, D.; Chen, L.; Liu, X.; Yan, X.; Tang, H.; and Xie, X. 2022. PPT: token-Pruned Pose Transformer for monocular and multi-view human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, 424–442. Springer.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7753–7762.

Qiu, H.; Wang, C.; Wang, J.; Wang, N.; and Zeng, W. 2019. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4342–4351.

Remelli, E.; Han, S.; Honari, S.; Fua, P.; and Wang, R. 2020. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6040–6049.

Shi, M.; Aberman, K.; Aristidou, A.; Komura, T.; Lischinski, D.; Cohen-Or, D.; and Chen, B. 2020. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40: 1–15.

Shuai, H.; Wu, L.; and Liu, Q. 2022. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Tan, S.; Zhen, X.; Xu, S.; Zheng, F.; He, Z.; and Shao, L. 2021. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210: 103225.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Xu, Q.; and Lin, S. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 507–523. Springer.

Zhang, J.; Cai, Y.; Yan, S.; Feng, J.; et al. 2021a. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34: 13153–13164.

Zhang, Z.; Wang, C.; Qiu, W.; Qin, W.; and Zeng, W. 2021b. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129: 703–718.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3425–3435.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.