Pattern Recognition 175 (2026) 113072



Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr





DSVTformer: Dual-stream spatial-view-temporal transformer for multi-view 3D human pose estimation

Wanruo Zhang [a], Mengyuan Liu [a,*], Wenhao Li [b], Hong Liu [a]

[a] State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, China
[b] Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Multi-View 3D human pose estimation
Spatial-view-temporal correlation modeling
Transformer

ABSTRACT

Current multi-view 3D human pose estimation methods typically use single-stream architectures that elevate 2D pose sequences into 3D space. However, their soley reliance on 2D joint coordinates limits performance by neglecting valuable complementary visual context from images. To address this, we propose a dual-stream approach that integrates both 2D pose and visual context, enhancing pose estimation without relying on camera parameters or body priors. Furthermore, existing transformer-based methods fail to adequately model the complex correlations across Spatial-View-Temporal (SVT) dimensions, focusing only on per-frame spatial modeling, local cross-view fusion, and temporal modeling over a single fused view. To overcome this, we introduce the Dual-stream Spatial-View-Temporal Transformer (DSVTformer) that fully captures SVT correlations, making it suitable for our dual-stream design. Its encoder extracts multi-view, multi-frame features from both image and pose modalities while a dual-stream decoder fuses them. Specifically, each decoder layer incorporates three axis-aware correlation blocks to model temporally enhanced spatial structures, global cross-view relations, and intra-view temporal dependencies. These blocks are grounded in a basic dual-stream interactive and enhancement mechanism, consisting of bidirectional cascaded cross-modal fusion modules and self-modal enhancement modules. This design allows DSVTformer to perform progressive, complementary cross-modal reasoning across spatial, inter-view, and temporal dimensions, significantly improving multi-view 3D human pose estimation. Extensive experiments on Human3.6M, MPI-INF-3DHP, and Ski-Pose demonstrate that DSVTformer achieves state-of-the-art performance in both accuracy and robustness across diverse multi-view settings. The code is available at https://github.com/Rowenazhang/DSVTformer.

1. Introduction

Estimating 3D human pose from visual data has become a fundamental task in computer vision due to its wide applications in action recognition, gait analysis, and human-object interaction. The objective is to accurately localize human joints in 3D space using monocular images, videos, or multi-view inputs. Current 3D human pose estimation (HPE) methods are commonly divided into single-stage and two-stage pipelines. Single-stage models [1,2] directly infer 3D poses from images, jointly learning feature extraction and spatial reasoning. While end-to-end, they suffer from high complexity and training instability. In contrast, two-stage methods [3] first detect 2D poses per view and then lift them to 3D, benefiting from strong 2D detectors and improved convergence. We focus on this two-stage paradigm.

Most two-stage methods rely on off-the-shelf 2D pose detectors and lift the resulting 2D joint coordinates to 3D space. However, the accuracy of 3D estimation is highly sensitive to the quality of the 2D inputs, which are often degraded by occlusion, motion blur, or background clutter. These errors propagate into the lifting stage, resulting in inaccurate or unstable 3D poses. To improve robustness against such unreliable 2D inputs, recent works have proposed uncertainty-aware estimation [4,5] and multi-hypothesis strategies [6,7] to handle detection ambiguity and missing joints. While effective to some extent, these methods rely solely on sparse 2D keypoints and overlook the rich visual context in the original images, limiting their ability to recover accurate and structurally consistent 3D poses in complex scenes. Moreover, they implicitly couple spatial, cross-view, and temporal correlations, without sequentially and explicitly disentangling them, limiting their multi-view spatiotemporal reasoning.

Some single-view 3D pose estimation methods integrate image features directly into the 2D-to-3D lifting process to enrich spatial context [11,12], while others incorporate temporal [13], spatiotemporal modeling [14,15], and human priors [16] to capture motion dynamics across frames. Despite these efforts, their reliance on a single RGB perspective

* Corresponding author.
E-mail addresses: wanruo_zhang@stu.pku.edu.cn (W. Zhang), liumengyuan@pku.edu.cn (M. Liu), wenhao.li@ntu.edu.sg (W. Li), hongliu@pku.edu.cn (H. Liu).

https://doi.org/10.1016/j.patcog.2026.113072
Received 16 August 2025; Received in revised form 4 January 2026; Accepted 8 January 2026
Available online 13 January 2026
0031-3203/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

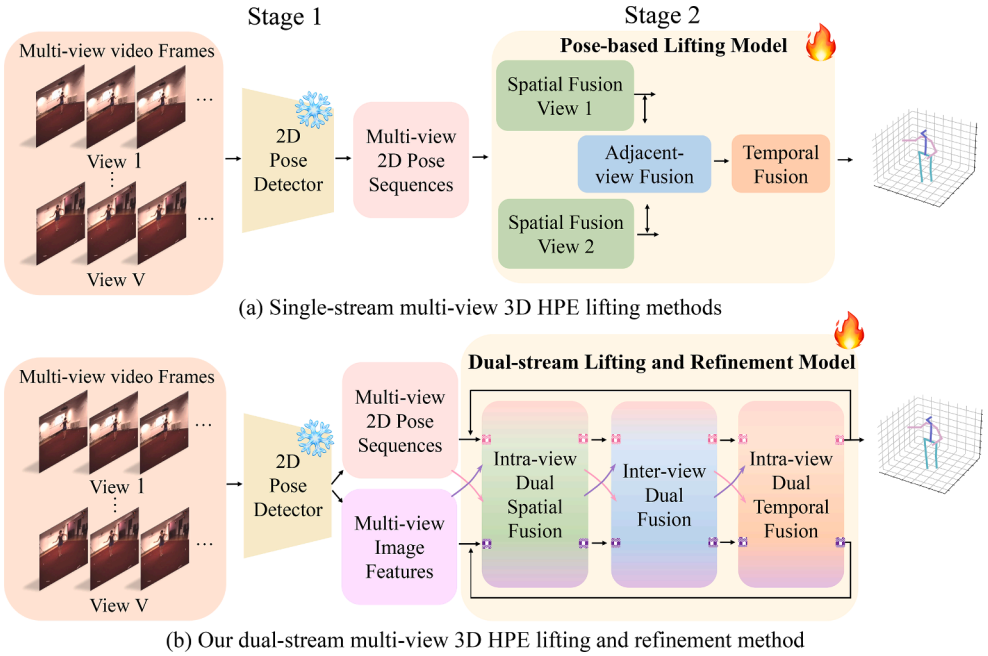


Fig. 1. (a) Previous methods [8–10] adopt a single-stream multi-view 3D HPE lifting approach, which freezes the 2D pose detector and then lifts 2D to 3D. (b) Our method introduces a dual-stream multi-view 3D HPE lifting and refinement framework that simultaneously utilizes 2D keypoints and rich image features. Interactive learning between the two modalities enables mutual enhancement of pose and visual cues, supporting fine-grained correlation modeling across spatial, cross-view, and temporal dimensions.

makes them inherently susceptible to depth ambiguity, self-occlusion, and limited perspective cues. In contrast, multi-view approaches leverage synchronized images from multiple cameras, providing richer geometric and spatial information. This enables more effective occlusion handling, depth disambiguation, and overall improvements in the stability and accuracy of 3D pose reconstruction, particularly in complex scenes.

Traditional multi-view 3D pose estimation methods often leverage epipolar geometry [17], triangulation [18], or explicit cross-view consistency constraints [19]. However, these approaches suffer from two inherent limitations. First, enforcing geometric constraints typically requires controlled environments and accurate camera calibration, which restricts their scalability and robustness in unconstrained real-world scenarios. Second, the strong reliance on predefined geometric assumptions and hand-crafted consistency rules may limit flexibility and generalization when facing severe occlusions, complex articulations, or heterogeneous camera configurations. Recent transformer-based lifting approaches [8–10] propose learning-based alternatives that implicitly model cross-view relationships without relying on explicit calibration. As shown in Fig. 1(a), these methods generally follow a single-stream lifting pipeline: (1) extract spatial relations within each view, (2) establish pairwise correspondences between adjacent views, and (3) fuse multi-view features followed by temporal modeling. Despite their effectiveness, several limitations remain as illustrated in Fig. 2(a). First, they overlook intra-view temporal dynamics, reducing sensitivity to motion cues. Second, the modeling of inter-view relationships is restricted to local view pairs, failing to capture global cross-view dependencies. Third, applying temporal modeling after multi-view fusion prevents the model from leveraging frame-specific view information. Lastly, these methods rely solely on 2D pose inputs without image context, which restricts 3D pose accuracy, occlusion handling, robustness to environmental variations, and generalization across diverse scenarios. These limitations hinder their performance in complex multi-camera environments.

To address the above limitations, we propose DSVTformer, a dual-stream spatial-view-temporal transformer for multi-view 3D HPE, as shown in Fig. 1(b). DSVTformer leverages multi-view, multi-frame images and 2D poses through three dual-stream fusion architectures, capturing complementary correlations across spatial, inter-view, and temporal dimensions to enhance 2D-to-3D lifting and improve reconstruction accuracy. Image context compensates for missing keypoints and provides semantic cues for globally consistent 3D poses, while pose information refines context by providing spatial relationships and motion cues. Specifically, DSVTformer consists of two stages: the first stage generates multi-view 2D keypoints and image context features using a 2D pose detector, and the second stage is a dual-stream lifting and refinement decoder, which processes and fuses pose and image features in parallel via a three-axis attention-based interactive and enhancement mechanism. This bidirectional design allows visual features to refine pose representations, while pose structures guide the enhancement of visual features.

To capture complex spatiotemporal dependencies in multi-view settings, we propose a sequential decomposition strategy with three specialized dual-stream blocks: Intra-view Dual Spatial (Intra-DS), Inter-view Dual (Inter-DV), and Intra-view Dual Temporal (Intra-DT) fusion blocks. These blocks are applied within each decoder layer to explicitly model dual-modal: (1) temporal-enhanced spatial relationships among joints within a single view (Intra-DS), (2) pairwise cross-view correlations among all camera views (Inter-DV), and (3) temporal dynamics across frames within each individual view (Intra-DT), as illustrated in Fig. 2(b). Each dual block operates through a Dual-stream Interactive and Enhancement Mechanism (DIEM), which consists of two cascaded cross-modal fusion modules, Image-to-Pose Interaction (I2PX) and Pose-to-Image Interaction (P2IX), along with a Self-modal Enhancement (SE) module. The I2PX module refines pose representations using visual context, while the P2IX module enhances visual representations based on pose-guided structural cues. This bidirectional interaction reduces reliance on 2D keypoints and improves robustness in pose recovery, while the SE module captures intra-modal dependencies for further refinement. By decoupling the multi-view spatiotemporal modeling along spatial, inter-view, and temporal axes, and capturing complementary cues between modalities at each axis, our model enables progressive and

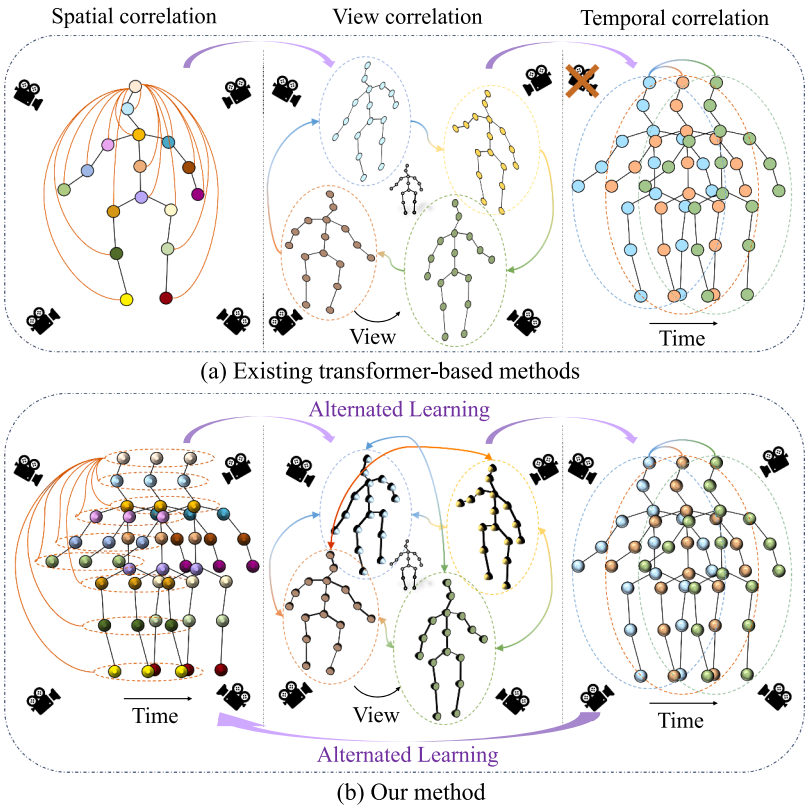


Fig. 2. Schematic illustrations of the receptive fields. (a) Existing methods [8–10] adopt a single-modality, pose-based architecture that performs multi-view feature fusion via single-frame spatial, adjacent view, and synthesized temporal fusion. (b) Our proposed method introduces a dual-stream architecture that integrates intra-view multi-frame spatial fusion, inter-view fusion, and intra-view temporal fusion across both image and pose features. Multi-layer alternating learning enables mutual compensation and progressive refinement between the two modalities. The cameras represent different viewpoints, while colored lines and circles indicate various modeling strategies. In (b), 3D volumetric joints represent fused pose-image features, whereas in (a), planar circles indicate joints modeled solely from pose data without visual context.

fine-grained reasoning across modalities. Through alternated stacking of multiple decoder layers, DSVTformer gradually enhances structural, cross-view, temporal, and modality-aware correlations, leading to improved robustness and accuracy in complex 3D HPE scenarios.

Our main contributions are summarized as follows:

- We propose DSVTformer, a transformer-based dual-stream framework that jointly processes image and pose modalities for multi-view 3D HPE. By integrating spatial, view, and temporal correlations along with visual context and pose cues, the model learns complementary cross-modal features, enhancing representation capacity and estimation accuracy.
- To address the limited expressiveness of 2D pose and underutilization of visual context, a dual-stream interactive and enhancement mechanism is designed with cascaded image-to-pose interaction and pose-to-image interaction modules for bidirectional cross-modal refinement, along with two self-modal enhancement modules to refine intra-modal features.
- A sequential decomposition strategy is proposed to tackle insufficient multi-view correlation modeling, where intra-view spatial, inter-view structural, and intra-view temporal dependencies are separately modeled using dedicated dual-stream fusion blocks. This axis-aware approach enables fine-grained, comprehensive multi-view integration.
- Extensive experiments demonstrate that DSVTformer achieves state-of-the-art performance in diverse scenes, without relying on camera calibration, human priors, or complex supervision, highlighting the effectiveness of cross-modal integration in multi-view 3D HPE.

2. Related works

2.1. Multi-view 3D HPE

Multi-view 3D human pose estimation leverages complementary observations from different camera perspectives to alleviate the inherent depth ambiguity and occlusion issues in single-view settings. Early methods addressed this ill-posed problem by incorporating strong structural priors, such as skeletal and bone constraints, and joint connectivity [20], but their reliance on hand-crafted designs limited generalization. To reduce this dependency, subsequent works introduced multi-view consistency as a supervisory signal. For example, Rhodin et al. [19] applied multi-view losses under weak supervision, while Pavlakos et al. [21] used geometric constraints to generate pseudo-labels from multi-camera 2D detections. Recent approaches have advanced multi-view 3D HPE by leveraging cross-view geometric relations through epipolar constraints [4,17,22–24], triangulation [25,26], or consensus-based optimization [27]. While effective, these methods typically rely on camera parameters or implicit calibration, leading to increased computational cost and reduced robustness under varying camera setups. To mitigate this, extrinsic-parameter-free methods such as [4,28–30] improve generalizability by integrating cross-view fusion, temporal modeling, and uncertainty-aware strategies. However, these methods still rely solely on 2D poses, underutilizing visual context and overlooking multi-dimensional correlations across spatial, temporal, and inter-view domains. In contrast, our transformer-based framework jointly processes image and pose features and explicitly decouples multi-view spatiotemporal reasoning along spatial, inter-view, and temporal axes,

enabling robust and fine-grained 3D pose estimation in uncalibrated settings.

2.2. Correlation modeling in transformer-based 3D HPE

Transformer architectures have become dominant in 3D HPE due to their effectiveness in modeling global dependencies. In single-view 3D HPE, various works [13,14] have explored different strategies based on transformer for modeling spatial, temporal, or spatiotemporal correlations with varying granularity and order. Some studies further integrate GCNs with transformers [3,31], enabling joint modeling of both local joint-level dependencies and global skeletal structures to improve single-view 3D pose estimation. More recently, multi-hop graph transformer architectures have been proposed to explicitly capture higher-order joint relations. For example, Islam et al. [32] introduce a multi-hop graph transformer network that combines graph-based message passing with transformer attention to model long-range dependencies across joints. While effective for enhancing structural reasoning in single-view settings, these approaches primarily focus on pose-only representations and do not address multi-view correlation modeling or the integration of visual context. However, single-view methods remain limited by occlusions and depth ambiguity. In comparison, multi-view methods provide richer observations from different perspectives and have become a major research direction for addressing these challenges more effectively in recent years. In the multi-view setting, transformer-based approaches require going beyond modeling of spatial and temporal correlations and explicitly capturing inter-view correlations. Methods such as [8–10] adopt staged pipelines, performing per-view per-frame spatial modeling, local cross-view fusion, and temporal modeling over the fused single-view features. However, these methods often constrain view fusion to local view pairs, neglect the spatial interactions among joints within temporally grouped sequences, or compromise temporal consistency in the original multi-view setting, leading to insufficient multi-view correlation modeling. Moreover, uncertainty-aware techniques have been introduced to improve robustness. For example, [4] integrates multi-view uncertainty via epipolar projection, while [30] incorporates 2D keypoint confidence into temporal transformers. Despite these efforts, existing methods remain heavily dependent on 2D keypoints, making them sensitive to detection errors and underutilizing visual context. While spatial-temporal and dual-stream transformers have shown promise in diverse tasks like infrared target detection [33], traffic prediction [34], and image enhancement [35,36], these methods primarily focus on homogeneous data or specific binocular settings. Consequently, they lack the capability to explicitly model the geometric consistency across multiple unconstrained views or fuse heterogeneous modalities (sparse pose and dense image features), which is addressed by our proposed triple-axis decomposition strategy. In contrast, our method does not depend on complex multi-view losses or camera parameters. Incorporating visual context into the pose through dual-stream fusion enhances spatial, view, and temporal correlations, achieving more accurate and robust multi-view 3D HPE.

2.3. Leveraging image features for 3D HPE

Lifting-based methods in 3D HPE, which directly regress 3D poses from 2D keypoints, are popular for their simplicity and efficiency. However, by neglecting image context, these methods are susceptible to errors from 2D detectors, especially under occlusion or noise. To mitigate this, recent works have explored integrating image features into 3D HPE pipelines. In single-view settings, Tekin et al. [37] combine global image features with heatmaps, while Zhao et al. [38] pool image features at joint locations and fuse them with 2D coordinates for graph-based reasoning. Han et al. [39] introduce uncertainty-aware scaling to reduce ambiguity, and Zhao et al. [11] utilize transformer models to learn joint representations from stacked image features and 2D pose maps. Zhang et al. [12] extend this to long video sequences via adaptive pooling for spatiotemporal feature extraction. However, most of these approaches are constrained to single-frame, single-view scenarios. In multi-view settings, image context can complement 2D pose information for robust cross-view reasoning, yet joint modeling of image-pose bimodality with spatial-view-temporal dependencies remains underexplored. We propose a dual-stream framework with attention-based interaction and refinement to address this. A sequential transformer captures dual-stream intra-view spatial, inter-view global, and intra-view temporal correlations, enabling unified multi-view spatiotemporal modeling.

3. Method

3.1. Overview

The proposed Dual-stream Spatial-View-Temporal Transformer (DSVTformer), shown in Fig. 3, comprises two stages: 2D feature extraction and 2D-to-3D lifting and refinement. The first stage extracts visual and pose features from multi-view, multi-frame images, while the second stage fuses these two complementary modalities for 3D HPE.

Given a sequence of multi-view RGB frames denoted as $\mathbf{F} \in \mathbb{R}^{T\times V\times H\times W\times 3}$, where T is the number of frames, V is the number of views, and H, W are the image height and width, we first perform human-centered preprocessing to obtain focused visual representations. Specifically, a pre-trained YOLOv3 detector [40] is employed to localize the human bounding box in each frame, which is used for image cropping and resizing to a fixed resolution. This step suppresses background clutter and ensures that subsequent feature extraction operates on compact, human-centric inputs. The cropped images are then processed by a pre-trained 2D pose detector, namely the Cascaded Pyramid Network (CPN) [41], which is used as a frozen feature extraction backbone. CPN is trained on large-scale human pose datasets with explicit keypoint supervision, enabling its intermediate representations to encode rich body-part and joint-related structural cues. In our implementation, we remove the heatmap prediction head and retain the ResNet–GlobalNet–RefineNet backbone as a frozen feature extractor. The highest-resolution feature map from the RefineNet module is selected, and global average pooling is applied to aggregate spatial context, resulting in compact per-frame-per-view image features $\mathbf{I} \in \mathbb{R}^{T\times V\times d}$, where d denotes the visual feature dimension. To facilitate interaction with the joint-indexed pose stream, each per-frame-per-view image feature is expanded into J tokens by repeating it along the joint axis, producing image tokens of shape $\mathbb{R}^{T\times V\times J\times d}$. This repetition serves as a tokenization and alignment preparation step that matches the token cardinality of the image stream with that of the joint-indexed pose stream. This repetition acts as a tokenization and alignment preparation step, providing a shared global visual context with matched token cardinality, such that each joint token can establish an explicit interaction channel with the image stream during the following dual-stream fusion. Meanwhile, the corresponding 2D pose features are represented as $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times C_{\text{in}}}$, where C_{in} denotes the dimension of the 2D keypoint coordinates.

In the second stage, the decoder is built upon the basic Dual-stream Interactive Enhancement Mechanism (DIEM), which consists of dual-stream bidirectionally cascaded cross-modal fusion modules for mutual interaction between pose and image modalities, followed by dual-stream Self-modal Enhancement (SE) for within-modality refinement. The decoder processes pose and image features through three sequential and alternating fusion blocks: Intra-view Dual Spatial (Intra-DS), Inter-view Dual (Inter-DV), and Intra-view Dual Temporal (Intra-DT), capturing spatial, view, and temporal correlations. By stacking multiple such spatial-view-temporal structures in an alternating manner, the final 3D pose is obtained by feeding the pose features refined in the last Intra-DT block into a regression head. This framework enables effective dual-modal integration across three correlation types, enhancing 2D-to-3D pose estimation in complex scenarios.

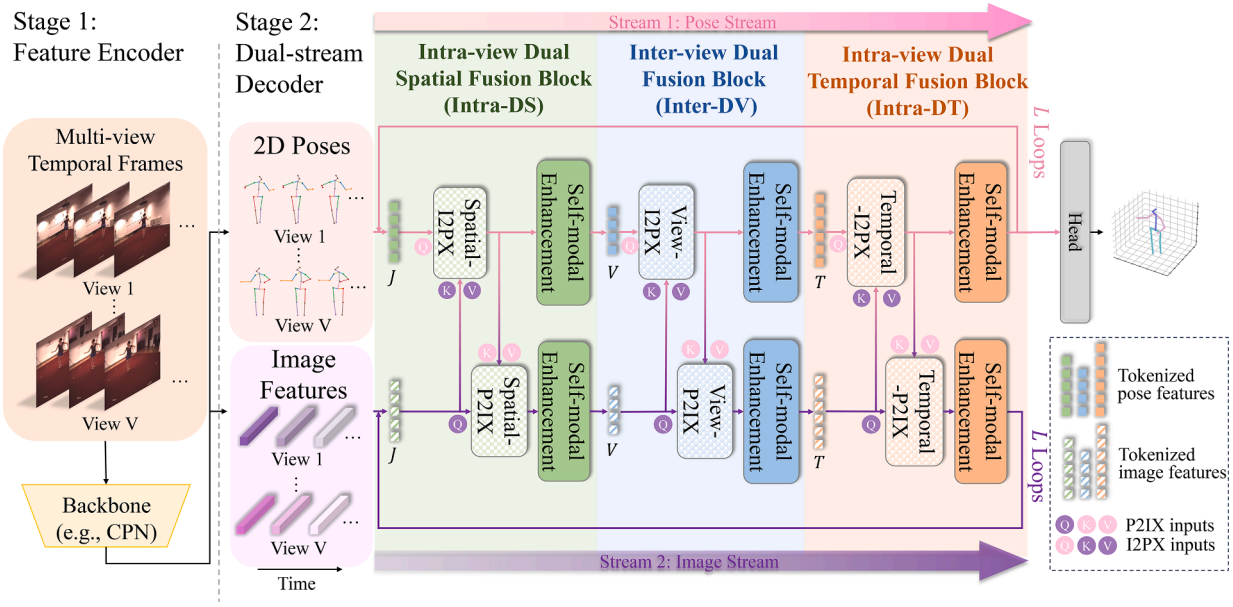


Fig. 3. Overview of the DSVTformer model. Using a backbone network, the first stage extracts 2D poses and image features from multi-view temporal image sequences. The second stage introduces a dual-stream decoder with multiple layers, each evolving through three sequential dual-stream fusion blocks: Intra-view Dual Spatial Fusion (Intra-DS), Inter-View Dual Fusion (Inter-DV), and Intra-view Dual Temporal Fusion (Intra-DT), enabling joint modeling of spatial, view, and temporal correlations, where J, V, and T denote the numbers of joint-, view-, and temporal tokens, respectively. Each block, based on the basic dual-stream interactive enhancement mechanism, combines cross-modal fusion through a bidirectional cascaded structure with Image-to-Pose Interaction (I2PX), followed by Pose-to-Image Interaction (P2IX), and Self-modal Enhancement (SE). The final 3D pose is regressed from the last layer's pose stream output.

3.2. Preliminary

Our model is built on the Transformer architecture to capture spatiotemporal dependencies of joint and visual features across views for 3D pose regression. Given an input sequence $\mathbf{X} \in \mathbb{R}^{n\times d}$, we briefly overview the core components: Multi-head Self-Attention (MSA), Multi-head Cross-Attention (MCA) and Multi-Layer Perceptron (MLP).

MSA. MSA models global dependencies by computing scaled dot-product attention in parallel across h heads. Scaled attention is defined as $\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}$, where $\mathbf{Q},\mathbf{K},\mathbf{V} \in \mathbb{R}^{n\times d_k}$ denote the query, key, and value matrices projected into the subspace of dimension d_k. For head i, $\text{head}_i = \text{Attention}(\mathbf{X}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{X}\mathbf{W}_i^{\mathbf{K}}, \mathbf{X}\mathbf{W}_i^{\mathbf{V}})$, with projection matrices $\mathbf{W}_i^{\mathbf{Q}}, \mathbf{W}_i^{\mathbf{K}}, \mathbf{W}_i^{\mathbf{V}} \in \mathbb{R}^{d\times d_k}$ and $d_k = d/h$. The h heads are concatenated and linearly projected as $\text{MSA}(\mathbf{X}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O$, where the output projection matrix $\mathbf{W}^O \in \mathbb{R}^{d\times d}$.

MCA. MCA operates similarly to MSA with multiple parallel attention heads, but instead of attending within a single sequence, it models the interactions between two distinct sequences. Here, the query is derived from the target sequence while the key and value come from the source sequence, enabling effective cross-modal or cross-view communication.

MLP. Following the attention module, the MLP (or FFN) module further transforms the features via two fully connected layers and a nonlinear activation function to enhance the model's representational capacity: $\text{MLP}(\mathbf{X}) = \sigma(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$, where $\sigma(\cdot)$ is the GELU activation function, $\mathbf{W}_1 \in \mathbb{R}^{d\times d_m}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_m\times d}$ are the weights, and $\mathbf{b}_1 \in \mathbb{R}^{d_m}$ and $\mathbf{b}_2 \in \mathbb{R}^d$ are the biases of the two layers.

3.3. Dual-stream decoder

3.3.1. Basic dual-stream interactive and enhancement mechanism

To effectively integrate 2D keypoints and image features for 3D human pose estimation (HPE), we propose a novel dual-stream decoder that exploits the complementary nature of pose and visual modalities. This decoder is built upon a fundamental Dual-stream Interactive and Enhancement Mechanism (DIEM), as illustrated in Fig. 4. Unlike prior approaches that either fuse modalities at a single stage or rely on one dominant modality, DIEM performs bidirectional, progressive dual-stream interaction and enhancement across three critical correlation axes: space, view, and time.

DIEM consists of two key components: dual-stream cross-modal fusion for cascaded bidirectional cross-modal interaction, and dual-stream Self-modal Enhancement (SE) for intra-modal feature refinement. Given input pose features $\mathbf{P}$ and image features $\mathbf{I}$, both are first tokenized along a correlation axis $a \in \{S, V, T\}$, representing Spatial, View, and Temporal dimensions respectively, yielding $\mathbf{P}_a$ and $\mathbf{I}_a$. After axis-specific projection and positional encoding, we obtain $\mathbf{X}_a^P$, $\mathbf{X}_a^I$. DIEM then applies dual cross-modal fusion followed by modality-specific self-enhancement.

The first step, Image-to-Pose Interaction (I2PX), enhances the pose stream by incorporating complementary context information from the image stream. Specifically, the image features $\mathbf{X}_a^I$ are first projected into the pose space and subsequently fused with the pose features $\mathbf{X}_a^P$ to generate context-guided pose representations $\mathbf{X}_a^{I\to P}$, as formulated below:

$$\begin{aligned}
&\mathbf{Q}_a^P = \mathbf{X}_a^P + \mathbf{E}_{Qa}^P, \mathbf{K}_a^I = \text{Proj}_{I\to P}(\mathbf{X}_a^I) + \mathbf{E}_{\mathcal{K}a}^P, \mathbf{V}_a^I = \mathbf{X}_a^I, \\
&\tilde{\mathbf{X}}_a^{I\to P} = \text{MCA}\big(\text{LN}(\mathbf{Q}_a^P),\ \text{LN}(\mathbf{K}_a^I),\ \text{LN}(\mathbf{V}_a^I)\big) + \mathbf{Q}_a^P, \\
&\mathbf{X}_a^{I\to P} = \text{MLP}(\text{LN}(\tilde{\mathbf{X}}_a^{I\to P})) + \tilde{\mathbf{X}}_a^{I\to P}.
\end{aligned} \tag{1}$$

Here, $\text{Proj}_{I\to P}$ denotes the projection of image features into the pose space; $\mathbf{E}_{Qa}^P$ and $\mathbf{E}_{\mathcal{K}a}^P$ are learnable pose-wise positional embeddings added to the query and key representations, respectively, to encode correlation-specific structure; $\mathbf{Q}_a^P$, $\mathbf{K}_a^I$, and $\mathbf{V}_a^I$ are the axis-specific query, key, and value representations, where the superscript P or I indicates that the feature originates from the pose or image stream, respectively. Subsequently, Pose-to-Image Interaction (P2IX) guides the image stream using the updated pose features, generating pose-aware visual representations. Following I2PX, the updated pose features $\mathbf{X}_a^{I\to P}$ serve as keys and values, while the image features $\mathbf{X}_a^I$ act as queries in P2IX, yielding

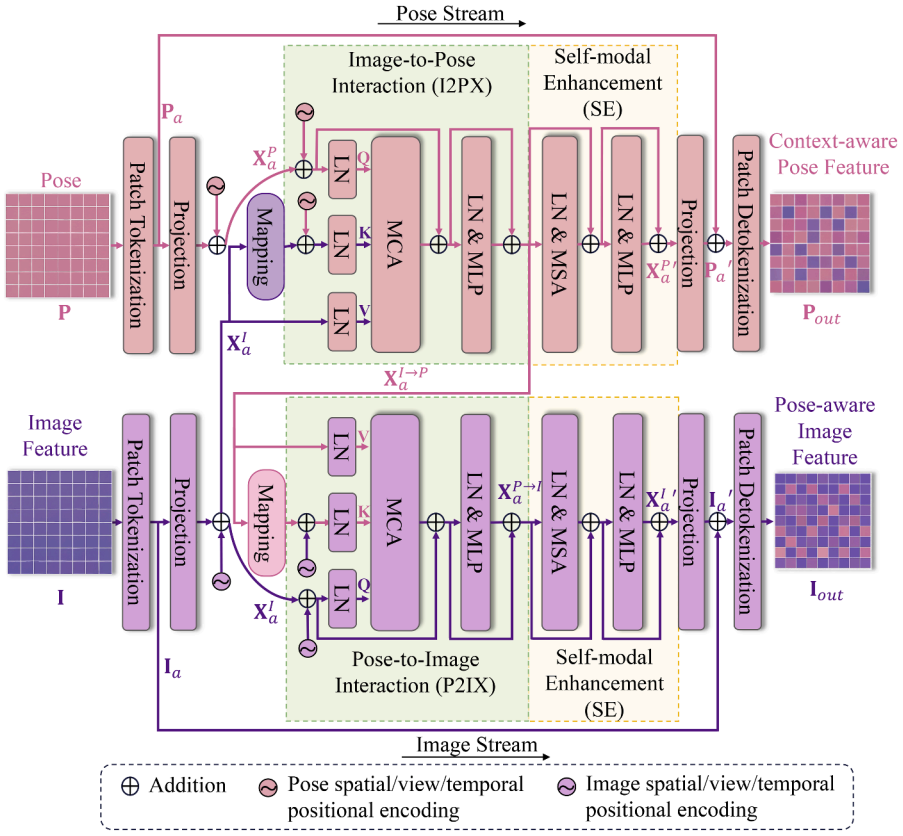


Fig. 4. The Dual-stream Interactive and Enhancement Mechanism (DIEM) enables bidirectional and cascaded cross-modal fusion between the pose stream (top) and the image stream (bottom) through Image-to-Pose Interaction (I2PX) and Pose-to-Image Interaction (P2IX), followed by Self-modal Enhancement (SE) in each stream. The subscript a denotes the correlation axis, which can be spatial, view, or temporal, in Intra-DS, Inter-DV, and Intra-DT, respectively.

the pose-guided image features $\mathbf{X}_a^{P\to I}$ computed as:

$$\begin{aligned}
\mathbf{Q}_a^I &= \mathbf{X}_a^I + \mathbf{E}_{\mathcal{Q}a}^I, \mathbf{K}_a^P = \mathrm{Proj}_{P\to I}(\mathbf{X}_a^{I\to P}) + \mathbf{E}_{\mathcal{K}a}^I, \mathbf{V}_a^P = \mathbf{X}_a^{I\to P},\\
\tilde{\mathbf{X}}_a^{P\to I} &= \mathrm{MCA}\big(\mathrm{LN}(\mathbf{Q}_a^I),\ \mathrm{LN}(\mathbf{K}_a^P),\ \mathrm{LN}(\mathbf{V}_a^P)\big) + \mathbf{Q}_a^I,\\
\mathbf{X}_a^{P\to I} &= \mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}}_a^{P\to I})) + \tilde{\mathbf{X}}_a^{P\to I},
\end{aligned} \tag{2}$$

where the projection function $\mathrm{Proj}_{P\to I}$ projects the context-guided pose representations $\mathbf{X}_a^{I\to P}$ back into the image space, facilitating the integration of pose information into the image stream, while $\mathbf{E}_{\mathcal{Q}a}^I$ and $\mathbf{E}_{\mathcal{K}a}^I$ are introduced as learnable context-wise positional embeddings applied to the query and key representations, respectively. While structurally symmetric, I2PX and P2IX serve distinct yet complementary roles: I2PX transfers contextual priors from the image stream to guide pose representation learning, whereas P2IX integrates structural cues from the pose stream to refine visual features. By operating with reversed and cascaded input roles, these modules enable progressive, bidirectional fusion, allowing each modality to be incrementally enriched with information from the other.

After cross-modal fusion, each stream undergoes independent refinement through Multi-head Self-Attention (MSA) and MLP, with layer normalization and residual connections. The following equations describe the Self-modal Enhancement (SE) for the pose and image streams, respectively. The context-guided pose features in the pose stream are refined as:

$$\begin{aligned}
\tilde{\mathbf{X}}_a^P &= \mathrm{MSA}(\mathrm{LN}(\mathbf{X}_a^{I\to P})) + \mathbf{X}_a^{I\to P},\\
\mathbf{X}_a^{P'} &= \mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}}_a^P)) + \tilde{\mathbf{X}}_a^P.
\end{aligned} \tag{3}$$

The pose-aware visual features in the image stream are refined as:

$$\begin{aligned}
\tilde{\mathbf{X}}_a^I &= \mathrm{MSA}(\mathrm{LN}(\mathbf{X}_a^{P\to I})) + \mathbf{X}_a^{P\to I},\\
\mathbf{X}_a^{I'} &= \mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}}_a^I)) + \tilde{\mathbf{X}}_a^I.
\end{aligned} \tag{4}$$

The refined pose and image representations are then projected back to their original embedding spaces and added to the corresponding input tokens via residual connections.

DIEM serves as a general core unit for dual-stream interaction and enhancement. In subsequent sections, we instantiate DIEM along three correlation axes: Intra-view Dual Spatial Fusion (Intra-DS) for spatial alignment within views, Inter-view Dual Fusion (Inter-DV) for geometric consistency across views, and Intra-view Dual Temporal Fusion (Intra-DT) for temporal coherence within views. These modules share a common DIEM structure but are sequentially adapted to operate along spatial, view, and temporal axes by customizing their interaction scope, tokenization strategy, and positional encoding, as detailed in later sections and illustrated in Algorithm 1.

3.3.2. *Intra-view dual spatial fusion block (Intra-DS)*

Our Intra-DS introduces a dual-stream spatial fusion strategy within each view, enabling fine-grained spatial-level alignment between pose and image tokens, reinforced by temporal information to effectively capture fine-grained cross-modal dependencies.

To capture intra-view spatial correlations, the Intra-view Dual Spatial Fusion Block (Intra-DS) instantiates the Dual-stream Interactive Enhancement Mechanism (DIEM) along the spatial axis. Given pose features $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times C_{S\ell}}$, where $C_{S\ell}$ is the pose channel dimension at layer ℓ, we reshape the input to treat views as independent instances within the effective batch dimension and flatten the temporal and channel axes into $C_T = C_{S\ell}\cdot T$. This yields pose tokens $\mathbf{P}_S \in \mathbb{R}^{J\times C_T}$, where joints J serve as the sequence length. Similarly, the extracted image features $\mathbf{I} \in \mathbb{R}^{T\times V\times J\times d}$ are reshaped into $\mathbf{I}_S \in \mathbb{R}^{J\times d_T}$, where $d_T = d\cdot T$. Both are linearly projected and added with learnable spatial positional embeddings $\mathbf{E}_S^P \in \mathbb{R}^{1\times J\times C_T'}$ and $\mathbf{E}_S^I \in \mathbb{R}^{1\times J\times d_T'}$ respectively, to obtain $\mathbf{X}_S^P \in \mathbb{R}^{J\times C_T'}$ and $\mathbf{X}_S^I \in \mathbb{R}^{J\times d_T'}$, with $C_T' = C'\cdot T$ and $d_T' = d'\cdot T$, where

Algorithm 1: Decoder Algorithm with DIEM.

Input : Axis: one of [Intra-DS, Inter-DV, Intra-DT], denoted by subscript a, 2D pose features: $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times C_{\text{in}}}$, Image features: $\mathbf{I} \in \mathbb{R}^{T\times V\times J\times d}$, Number of decoder layers: L, Layer index: l, Learnable positional embeddings for projected pose/image: $\mathbf{E}_a^P, \mathbf{E}_a^I$, Query, key, and value representations in MCA: $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$
(For clarity, the batch dimension is omitted throughout this formulation.)

Output: 3D pose estimation $\mathbf{Y}_{\text{3D}} \in \mathbb{R}^{T\times J\times C_{\text{out}}}$

for $l = 1$ **to** L **do**
  foreach axis $\in$ {Intra-DS, Inter-DV, Intra-DT} **do**
    // — Tokenization and Embedding —
    $(\mathbf{P}_a, \mathbf{I}_a) \leftarrow \text{Tokenize}(\mathbf{P}, \mathbf{I}, \text{axis})$
    $\mathbf{X}_a^P \leftarrow \text{PoseProj}(\mathbf{P}_a) + \mathbf{E}_a^P, \quad \mathbf{X}_a^I \leftarrow \text{ImageProj}(\mathbf{I}_a) + \mathbf{E}_a^I$
    // — Cross-modal Fusion —
    $\mathbf{Q}_a^P \leftarrow \mathbf{X}_a^P, \quad \mathbf{K}_a^I, \mathbf{V}_a^I \leftarrow \mathbf{X}_a^I$; // Prepare Q (pose), K/V (image) for I2PX
    $\mathbf{X}_a^{I\to P} \leftarrow \text{I2PX}(\mathbf{Q}_a^P, \mathbf{K}_a^I, \mathbf{V}_a^I)$; // Context-guided pose features
    $\mathbf{Q}_a^I \leftarrow \mathbf{X}_a^I, \quad \mathbf{K}_a^P, \mathbf{V}_a^P \leftarrow \mathbf{X}_a^{I\to P}$; // Prepare Q (image), K/V (pose) for P2IX
    $\mathbf{X}_a^{P\to I} \leftarrow \text{P2IX}(\mathbf{Q}_a^I, \mathbf{K}_a^P, \mathbf{V}_a^P)$; // Pose-guided image features
    // — Self-modal Enhancement —
    $\mathbf{X}_a^{P'} \leftarrow \text{SE}(\mathbf{X}_a^{I\to P}), \quad \mathbf{X}_a^{I'} \leftarrow \text{SE}(\mathbf{X}_a^{P\to I})$
    // — Residual Update and Detokenization —
    if $l > 1$ *or* axis $\neq$ Intra-DS **then**
      $\mathbf{P}'_a \leftarrow \text{Proj}(\mathbf{X}_a^{P'}) + \mathbf{P}_a, \quad \mathbf{I}'_a \leftarrow \text{Proj}(\mathbf{X}_a^{I'}) + \mathbf{I}_a$
    else
      $\mathbf{P}'_a \leftarrow \text{Proj}(\mathbf{X}_a^{P'}), \quad \mathbf{I}'_a \leftarrow \text{Proj}(\mathbf{X}_a^{I'})$
    end
    $(\mathbf{P}, \mathbf{I}) \leftarrow \text{Detokenize}(\mathbf{P}''_a, \mathbf{I}''_a, \text{axis})$
  end
end
$\mathbf{Y}_{\text{3D}} \leftarrow \text{RegressionHead}(\mathbf{P})$
return $\mathbf{Y}_{\text{3D}}$

C' and d' are the embedding dimension ratios. These embeddings explicitly encode joint identities to preserve spatial structure.

The spatial tokens $\mathbf{X}_S^P$ and $\mathbf{X}_S^I$, now enriched with temporal cues, are processed by DIEM for cross-modal interaction and self-modal enhancement. In this spatial correlation modeling stage, the Image-to-Pose Interaction module (Spatial-I2PX) attends to image tokens to extract contextual cues that help resolve pose estimation ambiguities, injecting them into the pose stream. Subsequently, the corresponding Pose-to-Image Interaction module (Spatial-P2IX) leverages the updated pose tokens to guide the image stream. By querying with pose tokens that possess explicit joint definitions, this interaction guides the image tokens to focus on joint-specific visual regions, progressively endowing the image stream-initially initialized with global features-with fine-grained joint semantics. Finally, the spatial-aware self-modal enhancement module applies self-attention and feedforward refinement within each modality for intra-stream enhancement.

The refined tokens are projected back to their embedding spaces and combined with residuals, yielding $\mathbf{P}''_S \in \mathbb{R}^{J\times(C_{out}\cdot T)}$ and $\mathbf{I}''_S \in \mathbb{R}^{J\times d_T}$, which are detokenized to match the original input format. As the first decoder layer lifting 2D to 3D, Intra-DS sets $C_{S1} = C_{in} = 2$ for $\ell = 1$ and $C_{S\ell} = C_{out} = 3$ for $\ell \geq 2$.

3.3.3. Inter-view dual fusion block (Inter-DV)

Following the intra-view spatial fusion in Intra-DS, the Inter-view Dual fusion block (Inter-DV) extends DIEM into the cross-view level, aiming to address viewpoint discrepancies and enforce geometric consistency across camera views. Unlike Intra-DS, which aligns pose and image features within a single view, Inter-DV explicitly models relationships between different views, enabling both modalities to exchange complementary geometric cues unavailable from a single viewpoint.

Given pose features $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times C_{V\ell}}$ with $C_{V\ell} = 3$, we apply a view-based tokenization strategy by folding the temporal dimension into the effective batch axis and flatten the joint and channel dimensions into $C_J = C_{V\ell} \cdot J$, yielding pose tokens $\mathbf{P}_V \in \mathbb{R}^{V\times C_J}$, where the number of views V serves as the sequence length. To preserve view identity without explicit calibration, we apply a linear projection to the tokens followed by the element-wise addition of learnable view-wise positional embeddings $\mathbf{E}_V^P \in \mathbb{R}^{1\times V\times C'_J}$ assigned to each view index, producing $\mathbf{X}_V^P \in \mathbb{R}^{V\times C'_J}$, where $C'_J = C' \cdot J$ and C' is the embedding ratio. Similarly, image features $\mathbf{I} \in \mathbb{R}^{T\times V\times J\times d}$ are tokenized into $\mathbf{I}_V \in \mathbb{R}^{V\times d_J}$, where $d_J = d \cdot J$, and projected into the latent space, augmented with distinct learnable view embeddings $\mathbf{E}_V^I \in \mathbb{R}^{1\times V\times d'_J}$ via summation, to yield $\mathbf{X}_V^I \in \mathbb{R}^{V\times d'_J}$, where $d'_J = d' \cdot J$. These embeddings allow the model to implicitly learn geometric dependencies from data without requiring manual relative ordering or camera extrinsics.

The view tokens $\mathbf{X}_V^P$ and $\mathbf{X}_V^I$ are processed by DIEM along the view axis. The View-aware Image-to-Pose Interaction module (View-I2PX) attends to image tokens from other viewpoints to extract complementary geometric cues for enhancing multi-view 3D reasoning in the pose stream. The corresponding Pose-to-Image Interaction module (View-P2IX) then uses the updated pose tokens to guide the image stream toward pose-consistent regions across views. The View-aware self-modal enhancement module applies view-wise self-attention and feedforward refinement within each modality, consolidating intra-modal coherence and improving robustness to viewpoint changes.

The refined tokens are projected back to their original embedding spaces and combined with residuals, producing $\mathbf{P}''_V \in \mathbb{R}^{V\times(C_{out}\cdot J)}$ and $\mathbf{I}''_V \in \mathbb{R}^{V\times d_J}$. These are then detokenized to match the original $\mathbf{P}$ and $\mathbf{I}$ shapes, ready for subsequent temporal modeling. By exploiting complementary cross-view cues, Inter-DV strengthens geometric consistency and reduces viewpoint ambiguity, complementing the spatial alignment achieved by Intra-DS.

3.3.4. Intra-view dual temporal fusion block (Intra-DT)

After the Inter-DV block aligns pose and image features across views, the Intra-view Dual Temporal fusion block (Intra-DT) focuses on temporal modeling within each view to capture motion dynamics and enhance frame-to-frame semantic consistency. It explicitly operates along the temporal axis, where each token corresponds to a single frame in a fixed view.

Given the pose features $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times C_{T\ell}}$ with $C_{T\ell} = 3$, distinct views are processed in parallel by incorporating the view axis into the effective batch dimension, and we flatten the joint–channel dimensions into $C_J = C_{T\ell} \cdot J$, yielding $\mathbf{P}_T \in \mathbb{R}^{T\times C_J}$, establishing a temporal sequence of length T. Similarly, image features $\mathbf{I} \in \mathbb{R}^{T\times V\times J\times d}$ are reshaped into $\mathbf{I}_T \in \mathbb{R}^{T\times d_J}$, where $d_J = d \cdot J$. Both are linearly projected and added with learnable temporal positional embeddings $\mathbf{E}_T^P \in \mathbb{R}^{1\times T\times C'_J}$ and $\mathbf{E}_T^I \in \mathbb{R}^{1\times T\times d'_J}$ to encode frame order, obtaining $\mathbf{X}_T^P \in \mathbb{R}^{T\times C'_J}$ and $\mathbf{X}_T^I \in \mathbb{R}^{T\times d'_J}$.

DIEM then processes the temporal tokens along the time axis. The Temporal-aware Image-to-Pose Interaction module (Temporal-I2PX) attends to image tokens across frames to extract temporally evolving contextual information, injecting it into the pose stream to enhance temporal coherence in pose estimation. Correspondingly, Temporal-P2IX leverages the updated temporal pose tokens to guide the image stream, focusing visual representations on pose-relevant movements. Temporal-aware self-modal enhancement module applies frame-wise self-attention and feedforward refinement within each modality to strengthen intra-stream temporal coherence.

The refined outputs $\mathbf{X}_T^{P'}$ and $\mathbf{X}_T^{I'}$ are projected back to their original embedding spaces and combined with residuals, producing $\mathbf{P}''_T \in$

$\mathbb{R}^{T\times C_{T\ell}\cdot J}$ and $\mathbf{I}''_T \in \mathbb{R}^{T\times d_J}$. These are detokenized to match the original $\mathbf{P}$ and $\mathbf{I}$ shapes. If $\ell = L$, the pose output is fed into the regression head for final 3D pose estimation; otherwise, they are passed to the next decoder layer.

4. Experiments and results

In this section, we present comprehensive experimental evaluations to validate the effectiveness of our approach. We begin by describing the datasets and evaluation metrics used in our study, followed by implementation details that ensure reproducibility. Subsequently, we perform quantitative comparisons with state-of-the-art methods to highlight the superiority of our model. In addition, we provide qualitative results to illustrate the quality of the predictions. Finally, we conduct a series of ablation studies to investigate the contribution of each component within our framework.

4.1. Datasets and metrics

We evaluate the effectiveness and generalization ability of our proposed method on three widely used 3D human pose estimation benchmarks: Human3.6M [42], MPI-INF-3DHP [43], and Ski-Pose PTZ-Camera [44]. These datasets encompass controlled indoor environments, mixed indoor-outdoor settings, and challenging in-the-wild scenes, providing a comprehensive assessment across various scenarios.

Human3.6M. The Human3.6M dataset [42] is a large-scale benchmark for 3D HPE, containing about 3.6 million frames of 11 subjects performing 15 everyday activities captured from multiple cameras in a controlled indoor setting. Following common practice, we train on subjects S1, S5, S6, S7, and S8, and evaluate on S9 and S11. Performance is reported with two standard metrics: mean per-joint position error (MPJPE, P1), which directly computes the average Euclidean distance between predicted and ground-truth joint coordinates, and Procrustes-aligned MPJPE (PA-MPJPE, P2), which measures the MPJPE after applying a rigid alignment through Procrustes analysis.

MPI-INF-3DHP. The MPI-INF-3DHP dataset [43] provides a diverse benchmark for 3D HPE, with over one million frames of 8 subjects captured from 14 cameras in both indoor and outdoor scenes. Its test set covers three environments-green-screen studio, non-green-screen studio, and outdoor-making it suitable for evaluating robustness and generalization. We report performance using four widely accepted metrics: MPJPE (P1), PA-MPJPE (P2), Percentage of Correct Keypoints (PCK) with a threshold of 150 *mm*, and the Area Under the PCK Curve (AUC).

Ski-Pose PTZ-Camera. The Ski-Pose PTZ-Camera dataset [44] introduces a challenging in-the-wild scenario, capturing alpine skiers during giant slalom events in snowy outdoor environments. It features six skiers recorded from six different viewpoints using pan-tilt-zoom cameras. The dataset is specifically designed to test pose estimation methods in dynamic and unconstrained conditions. Following the official split, we use data from subjects 1 through 5 for training and subject 6 for evaluation. The performance on this dataset is assessed using MPJPE and PA-MPJPE (P1 and P2), enabling a consistent evaluation framework across different datasets.

4.2. Implementation details

All experiments are conducted on a single NVIDIA RTX 4090 GPU. We adopt different encoder settings depending on the dataset; by default, a pretrained CPN [41] is used as both the 2D pose estimator and the image feature extractor. For 2D pose inputs, both ground-truth and CPN-estimated keypoints are used on Human3.6M [42], whereas ground-truth 2D poses are adopted for MPI-INF-3DHP [43] and Ski-Pose [19]. The proposed DSVTformer is trained for 50 epochs using the Adam optimizer [45] with an initial learning rate of 0.0002, decayed by a factor of 0.98 per epoch. The dual-stream decoder consists of $L = 2$ layers, with image and pose embedding coefficients set to $d' = 16$ and $C' = 32$, respectively. No camera parameters are required during training or inference, and an L2 loss is used for supervision.

4.3. Comparison with state-of-the-art methods

Results on Human3.6M. Table 1 reports a quantitative comparison on the Human3.6M dataset under the P1 protocol. Our proposed method achieves an average MPJPE of 25.4 *mm*, outperforming all single-view approaches, including the previous best-performing method [16] (33.0 *mm*), demonstrating that the multi-view multi-frame design effectively alleviates the depth ambiguity problem commonly observed in monocular settings. Compared to multi-view methods that rely on known camera parameters, the performance of our DSVTformer is highly competitive, slightly trailing [25] (20.8 *mm*), while offering greater flexibility in uncalibrated environments. Among methods not relying on camera calibration, our approach achieves state-of-the-art performance, outperforming the most recent method [10] (27.4 *mm*) and the previously leading approach [4] (26.4 *mm*) by 2.0 *mm* and 1.0 *mm*, respectively. Notably, ours yields the best results in 12 out of 15 actions, and ranks second in the remaining 3, indicating consistent effectiveness across diverse human poses. These results suggest that jointly modeling 2D poses and image context enables richer semantic representation and robust correlation learning, while reducing the impact of inaccurate 2D poses. When ground-truth 2D keypoints are used, the model achieves 12.5 *mm* MPJPE, substantially better than 25.4 *mm* with CPN-detected poses, highlighting the importance of 2D accuracy and the model's ability to exploit precise inputs. Compared with other strong GT-based methods such as [30] (15.3 *mm*) and [10] (17.6 *mm*), the proposed model achieves the lowest error. All results are obtained without extra data or supervision, using a simple L2 loss, confirming the method's efficiency and scalability.

Results on 3DHP. Table 2 summarizes results on the MPI-INF-3DHP dataset under two settings: (1) trained and tested on 3DHP, and (2) trained on Human3.6M and fine-tuned on 3DHP. Under the first setting, the proposed method achieves a P1 error of 9.9 *mm*, surpassing the most recent method [10] (16.9 *mm*) and the previously strongest method [9] (12.1 *mm*) by 7.0 *mm* and 2.2 *mm*, respectively. All other metrics, including PCK (99.9) and AUC (93.0), are also the best among the compared methods. In the cross-dataset setting, the model achieves even better results, with a P1 of 8.6 *mm* and P2 of 5.7 *mm*, outperforming all prior works. Notably, the cross-dataset setting yields even better performance than the in-domain setting, indicating strong generalization and cross -dataset adaptability.

Results on Ski-Pose. Table 3 presents results on the Ski-Pose dataset under two training settings: (1) trained and tested on Ski-Pose, and (2) pretrained on Human3.6M then fine-tuned on Ski-Pose. In the first setting, the proposed method achieves a P1 of 60.9 *mm* and a P2 of 42.6 *mm*, improving over the best previous method [10] by 0.3 *mm* and 3.2 *mm*, respectively. In the second setting, the model achieves further improvements, with a P1 of 40.1 *mm* and P2 of 27.1 *mm*, surpassing all existing methods, including [10], by 7.8 *mm* and 6.5 *mm*, respectively. Notably, the substantial performance gains in the cross-dataset setting, reducing P1 by 20.8 *mm* and P2 by 15.5 *mm* compared to the in-domain baseline, demonstrate the effectiveness of large-scale indoor pretraining and highlight the model's strong generalization ability to challenging in-the-wild sports scenarios.

4.4. Ablation studies

Extensive ablation studies are conducted on the Human3.6M dataset with 2D detected inputs to investigate the impact of different design choices in our model, using the default input setting of 27 video frames.

Table 1
Comparison with state-of-the-art 3D HPE methods on the Human3.6M dataset (P1, *mm*). The best results are in *italic*, and the second-best are underlined.

Methods	Venue	2D Pose	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Single-view methods																		
Chen [46]	TCSVT'21	(CPN, T=243)	41.4	43.2	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
Zhang [14]	CVPR'22	(CPN, T=243)	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
Ci [47]	CVPR'23	(SH, T=1)	31.7	35.4	31.7	32.3	36.4	42.4	**32.7**	31.5	41.2	52.7	36.5	34.0	36.2	29.5	30.2	35.6
Li [48]	PR'23	(CPN, T=351)	39.1	42.7	38.7	40.3	44.1	50.0	41.4	38.7	53.9	61.6	43.6	40.8	42.5	29.6	30.6	42.5
Chen [3]	PR'24	(CPN, T=1)	46.3	50.3	45.7	50.5	50.8	57.5	49.6	46.0	55.8	63.8	50.9	47.8	53.0	38.7	41.3	49.8
Zhou [49]	AAAI'24	(CPN, T=1)	44.9	46.4	42.4	44.9	48.7	40.1	44.3	55.0	58.9	47.1	48.2	42.6	36.9	48.8	40.1	46.4
Du [15]	PR'24	(CPN, T=243)	36.1	40.3	41.9	33.3	35.7	61.0	41.0	33.6	**37.4**	69.9	38.4	36.7	43.4	25.6	29.4	40.2
Zhao [50]	NeurIPS'24	(HRNet, T=1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39.8
Zhang [12]	MM'24	(HRNet, T=243)	33.7	37.2	35.0	32.2	37.2	40.7	34.6	34.3	45.1	51.4	37.0	34.6	36.8	27.0	26.8	36.2
Peng [16]	CVPR'24	(CPN, T=243)	**30.1**	**32.1**	**29.1**	**30.6**	**35.4**	**39.3**	32.8	**30.9**	43.1	**45.5**	**34.7**	**33.2**	**32.7**	**22.1**	**23.0**	**33.0**
Multi-view methods (camera parameters are **given**)																		
Iskakov [25]	ICCV'19	(SB, T=1)	**19.9**	**20.0**	**18.9**	**18.5**	**20.5**	**19.4**	**18.4**	**22.1**	**22.5**	28.7	**21.2**	**20.8**	**19.7**	**22.1**	**20.2**	**20.8**
Qiu [22]	ICCV'19	(HRNet, T=1)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	**26.9**	31.0	25.6	25.0	28.1	24.4	26.2
He [23]	CVPR'20	(SH, T=1)	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	26.9
Bultmann [24]	RSS'21	(SB, T=1)	27.1	29.9	27.0	26.5	31.3	28.9	27.1	29.8	36.5	36.0	30.8	29.3	29.7	27.3	26.3	29.8
Luvizon [27]	IJCV'22	(GT, T=1)	31.0	33.0	41.0	34.0	41.0	37.0	37.0	51.0	56.0	43.0	44.0	37.0	33.0	42.0	32.0	39.0
Bartol [26]	CVPR'22	(SB, T=1)	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1
Shuai [30]	TPAMI'22	(CPN, T=27)	23.4	25.2	23.1	24.4	27.4	28.5	22.8	25.2	28.7	36.2	25.9	23.6	26.6	22.6	22.7	25.8
Davoodnia [4]	ECCV'24	(CPN, T=27)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26.2
Multi-view methods (camera parameters are **not given**)																		
Huang [29]	WACV'20	(SH, T=1)	26.8	32.0	25.6	52.1	33.3	42.3	25.8	25.9	40.5	76.6	39.1	54.5	35.9	25.1	24.2	37.5
Remelli [18]	CVPR'20	(-, T=1)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	**31.7**	31.2	29.9	26.9	33.7	30.4	30.2
Shuai [30]	TPAMI'22	(CPN, T=27)	24.6	**25.4**	24.8	24.6	28.7	29.1	23.9	25.6	31.4	36.2	26.6	24.7	28.9	23.7	23.6	26.6
Zhou [8]	MM'23	(CPN, T=27)	24.8	27.7	24.3	24.9	27.7	29.8	24.5	25.3	30.5	33.4	28.2	24.0	28.4	24.7	24.3	26.8
Davoodnia [4]	ECCV'24	(CPN, T=27)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26.4
Zhang [9]	AAAI'24	(CPN, T=27)	26.5	28.3	23.0	25.9	27.2	31.0	25.4	27.2	**28.6**	33.8	28.6	25.6	30.1	27.1	26.5	27.6
Zhang [10]	PR'25	(CPN, T=27)	**24.0**	27.5	24.8	25.0	28.8	31.3	24.7	27.0	30.6	36.2	28.6	25.3	30.0	24.7	24.0	27.4
Ours		(CPN, T=27)	24.1	26.7	**22.3**	**23.3**	**26.0**	**27.8**	**23.4**	**24.3**	**28.6**	32.0	**26.1**	**22.6**	**26.8**	**23.3**	**23.5**	**25.4**
Luvizon [27]	IJCV'22	(GT, T=1)	40.0	36.0	44.0	39.0	44.0	42.0	41.0	66.0	70.0	46.0	49.0	43.0	34.0	46.0	34.0	45.0
Gordon [28]	ECCV'22	(GT, T=27)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22.9
Shuai [30]	TPAMI'22	(GT, T=27)	15.5	17.1	13.7	15.5	14.0	16.2	15.8	16.5	15.8	16.1	14.5	14.5	16.9	14.3	13.7	15.3
Zhang [10]	PR'25	(GT, T=27)	16.5	18.3	15.1	17.2	17.0	19.4	16.8	18.7	16.4	22.3	17.5	17.3	20.0	15.4	16.8	17.6
Ours		(GT, T=27)	**12.7**	**13.7**	**10.9**	**12.3**	**11.1**	**15.3**	**12.3**	**13.2**	**11.6**	**14.2**	**12.1**	**11.4**	**14.4**	**10.9**	**11.5**	**12.5**

Table 2
Comparison results on MPI-INF-3DHP dataset. T is the number of input frames. The best results are in *italic*, and the second-best results are underlined.

Methods	Venue	Trainset	T	PCK ↑	AUC ↑	P1 (*mm*) ↓	P2 (*mm*) ↓
Kocabas [17]	CVPR'19	3DHP	1	77.5	-	109.0	-
Wandt [51]	CVPR'21	3DHP	1	77.0	-	104.0	70.3
Gholami [52]	Neurocomput'22	3DHP	27	-	-	101.5	76.5
Zhou [8]	MM'23	3DHP	27	98.7	86.8	18.0	13.1
Zhang [9]	AAAI'24	3DHP	27	98.7	90.2	16.9	12.1
Zhang [10]	PR'25	3DHP	27	99.5	74.8	21.8	16.7
Ours		3DHP	27	**99.9**	**93.0**	**9.9**	**6.8**
Zhou [8]	MM'23	H36M	27	99.8	88.5	15.5	11.4
Zhang [9]	AAAI'24	H36M	27	**99.9**	91.7	10.6	7.6
Zhang [10]	PR'25	H36M	27	99.8	75.2	18.6	14.8
Ours		H36M	27	**99.9**	**93.9**	**8.6**	**5.7**

Variants of dual-stream spatial-view-temporal transformer. Fig. 5 shows five variants of the dual-stream spatial-view-temporal transformer. (a) is the proposed sequentially alternating model (DSVT-former), which models spatial, view, and temporal correlations in a strictly ordered and alternating manner. (b) adopts a plain sequential design without interleaving across stages. (c) performs all three correlations in parallel within each layer and integrates them via an Adaptive Fusion (AF) module that applies global max pooling along each dimension, followed by concatenation and attention-based weighting. (d) and (e) are hybrid variants that combine sequential and parallel modeling in different orders. As shown in Table 4, model (a) achieves the best performance (P1: 25.4 *mm*, P2: 19.9 *mm*), surpassing model (b), which lacks the alternating design, by 1.1 *mm* and 0.5 *mm*, respectively. This demonstrates the effectiveness of alternating correlation modeling. Moreover, model (a) consistently outperforms other sequence-aware variants such as (c), (d), and (e), confirming that the sequentially alternating design is superior not only to parallel-only strategies, but also to hybrid sequential-and-parallel and non-alternating fusion schemes, as it gradually integrates dual-modal spatial, view, and temporal information in a structured order.

Correlation order for dual-stream fusion blocks. To assess the impact of correlation ordering in the dual-stream fusion pipeline, we evaluate all six permutations of the spatial ($\mathbb{S}$), view ($\mathbb{V}$), and temporal ($\mathbb{T}$) blocks. As shown in Table 5, model performance is sensitive to the sequence of these modules. The best results are obtained with the $\mathbb{S} \rightarrow \mathbb{V} \rightarrow \mathbb{T}$ order, which is adopted in our final design. In contrast, placing the spatial block later-e.g., $\mathbb{T} \rightarrow \mathbb{V} \rightarrow \mathbb{S}$ or $\mathbb{V} \rightarrow \mathbb{T} \rightarrow \mathbb{S}$-leads to a performance drop of over 4 *mm* in P1. These findings underscore

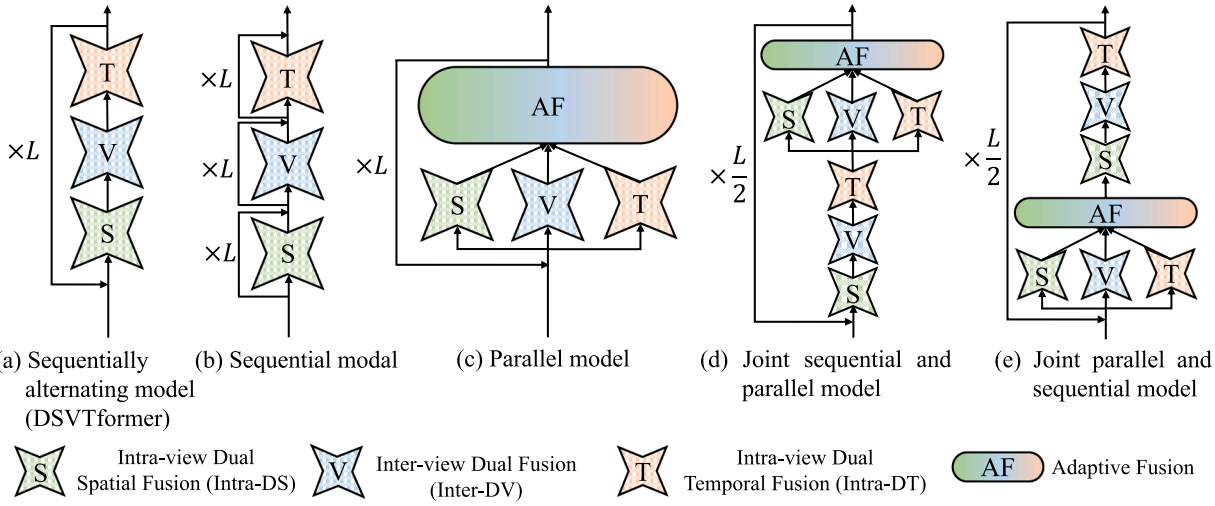


Fig. 5. Variants of the dual-stream spatial-view-temporal transformer. The proposed (a) sequentially alternating model (DSVTformer) yields the best performance.

Table 3
Comparisons on Ski-Pose dataset. T is the number of input frames. The best results are in *italic*, and the second-best results are underlined.

Methods	Venue	Trainset	T	P1 (*mm*) ↓	P2 (*mm*)↓
Rhodin [19]	CVPR'18	Ski	1	85.0	-
Wandt [51]	CVPR'21	Ski	1	128.1	89.6
Gordon [28]	ECCV'22	Ski	27	65.5	-
Zhou [8]	MM'23	Ski	27	62.6	49.4
Zhang [9]	AAAI'24	Ski	27	63.2	48.5
Zhang [10]	PR'25	Ski	27	<u>61.2</u>	<u>45.8</u>
Ours		Ski	27	**60.9**	**42.6**
Zhou [8]	MM'23	H36M	27	<u>42.2</u>	<u>29.4</u>
Zhang [9]	AAAI'24	H36M	27	45.3	31.4
Zhang [10]	PR'25	H36M	27	47.9	33.6
Ours		H36M	27	**40.1**	**27.1**

Table 4
Ablation study on the variants of dual-stream spatial-view-temporal transformer.

ID	Model	P1 (*mm*) ↓	P2 (*mm*) ↓
(a)	Sequentially alternating model (DSVTformer)	**25.4**	**19.9**
(b)	Sequential modal	26.5	20.4
(c)	Parallel weighted model	29.3	22.7
(d)	Joint sequential and parallel model	25.9	20.1
(e)	Joint parallel and sequential model	29.2	22.9

Table 5
Ablation study on sequential correlation order for dual-stream fusion blocks. $\mathbb{S}, \mathbb{V}, \mathbb{T}$ represent Intra-DS, Inter-DV, and Intra-DT blocks, respectively. Placing spatial modeling first (*i.e.*, $\mathbb{S} \rightarrow \mathbb{V} \rightarrow \mathbb{T}$) yields the best performance.

Sequential correlation orders	P1(*mm*)↓	P2(*mm*)↓
$\mathbb{V} \rightarrow \mathbb{S} \rightarrow \mathbb{T}$	28.7	22.7
$\mathbb{V} \rightarrow \mathbb{T} \rightarrow \mathbb{S}$	29.5	23.4
$\mathbb{T} \rightarrow \mathbb{V} \rightarrow \mathbb{S}$	29.5	23.5
$\mathbb{T} \rightarrow \mathbb{S} \rightarrow \mathbb{V}$	28.9	22.9
$\mathbb{S} \rightarrow \mathbb{T} \rightarrow \mathbb{V}$	26.5	20.3
$\mathbb{S} \rightarrow \mathbb{V} \rightarrow \mathbb{T}$	**25.4**	**19.9**

the importance of introducing pose-image spatial correlations early to guide subsequent view and temporal interactions effectively.

Model components. We conduct comprehensive ablation studies on Human3.6M to analyze the impact of individual components in DSVTformer (Table 6). Removing the image stream, self-modal enhancement, or cross-modal interaction consistently leads to noticeable accuracy degradation, confirming that visual context and dual-stream interactions provide essential complementary information beyond pose-only modeling. Among the interaction variants, removing the image-to-pose interaction causes the largest performance drop, indicating its critical role in guiding pose representations with visual cues. For spatial-view-temporal modeling, removing spatial fusion results in the most severe accuracy degradation, while removing view or temporal fusion leads to comparatively smaller but still consistent performance losses. Variants that retain only partial correlations (e.g., view-only or temporal-only modeling) show substantially inferior accuracy despite significantly reduced model size and computation, suggesting that effective 3D pose estimation requires joint reasoning across spatial, view, and temporal dimensions. Overall, the full DSVTformer achieves the best pose estimation accuracy under practical real-time constraints, attaining 25.4/19.9 *mm* on P1/P2 with a runtime of 73 FPS, and outperforming all ablated variants that meet real-time inference requirements. These results indicate that the proposed architecture delivers superior accuracy while satisfying real-time deployment constraints in multi-view 3D human pose estimation.

Variants of dual-stream fusion strategies. We conduct ablation studies to evaluate different dual-stream fusion strategies, as illustrated in Fig. 6 and summarized in Table 7. The best performance is achieved by two streams interact with KV configuration (i), used in DSVTformer. Removing all interactions (a) results in a substantial degradation (P1 increases by 10.5 *mm*), highlighting the importance of dual-stream interaction. Using only one Q/K/V from the other stream (d-f) yields worse results, indicating that partial interaction is insufficient. Moreover, combining two inputs from the complementary stream (g, h), such as using QK or QV, still underperforms compared to KV fusion. These findings suggest that simultaneously using K and V from the other modality provides the most balanced and effective cross-modal fusion.

Input 2D detectors and stream settings. To evaluate the robustness of our approach under different 2D detection qualities and to analyze the contribution of the image context stream, we conduct an ablation study on Human3.6M using 2D pose inputs generated by three widely adopted off-the-shelf detectors: Cascaded Pyramid Network (CPN) [41], Stacked Hourglass (SH) [53], and HRNet [54]. The quantitative results are reported in Table 8. When the full dual-stream architecture is employed, replacing the default CPN detector with SH or HRNet leads to only moderate performance variations, with P1 ranging from 25.4 to 27.0 *mm* and P2 from 19.9 to 20.7 *mm*. Despite noticeable differences in detector architectures and localization accuracy, the overall performance degradation remains limited, indicating that the proposed method maintains stable performance across different 2D input qualities. Notably, the

Table 6
Ablation study on different components in DSVTformer, reporting pose estimation performance, model complexity (Params and FLOPs), and runtime efficiency (FPS). The full model achieves the best accuracy among all variants that satisfy real-time inference requirements.

ID	Method	Pose stream		Image stream		DIEM			P1 (*mm*)↓	P2 (*mm*)↓	Params (M)	FLOPs (G)	FPS
		I2PX	SE	P2IX	SE	Intra-DS	Inter-DV	Intra-DT					
(a)	DSVTformer (ours)	✓	✓	✓	✓	✓	✓	✓	**25.4**	**19.9**	104.72	8.61	73
(b)	w/o image context	✓	✓	✓	✗	✓	✓	✓	27.6	21.3	67.72	5.81	139
(c)	w/o pose self-modal enhancement	✓	✗	✓	✓	✓	✓	✓	27.6	21.5	72.54	5.84	92
(d)	w/o context self-modal enhancement	✓	✓	✓	✗	✓	✓	✓	25.6	20.1	96.66	8.01	86
(e)	w/o both self-modal enhancement	✓	✗	✓	✗	✓	✓	✓	27.4	21.1	64.49	5.24	112
(f)	w/o image-to-pose interaction	✗	✓	✓	✓	✓	✓	✓	28.0	22.1	74.54	5.88	108
(g)	w/o pose-to-image interaction	✓	✓	✗	✓	✓	✓	✓	26.0	20.3	92.63	7.71	95
(h)	w/o both interactions	✗	✓	✗	✓	✓	✓	✓	27.9	21.8	62.45	4.99	136
(i)	w/o spatial fusion	✓	✓	✓	✓	✗	✓	✓	32.5	26.2	46.37	4.64	101
(j)	w/o view fusion	✓	✓	✓	✓	✓	✗	✓	27.8	21.1	81.59	6.11	97
(k)	w/o temporal fusion	✓	✓	✓	✓	✓	✓	✗	26.1	20.1	81.48	6.10	115
(l)	w/o view and temporal fusion	✓	✓	✓	✓	✓	✗	✗	28.8	22.2	58.35	3.38	190
(m)	w/o spatial and temporal fusion	✓	✓	✓	✓	✗	✓	✗	28.4	22.1	23.13	2.13	205
(n)	w/o spatial and view fusion	✓	✓	✓	✓	✗	✗	✓	28.0	21.3	23.24	2.14	183

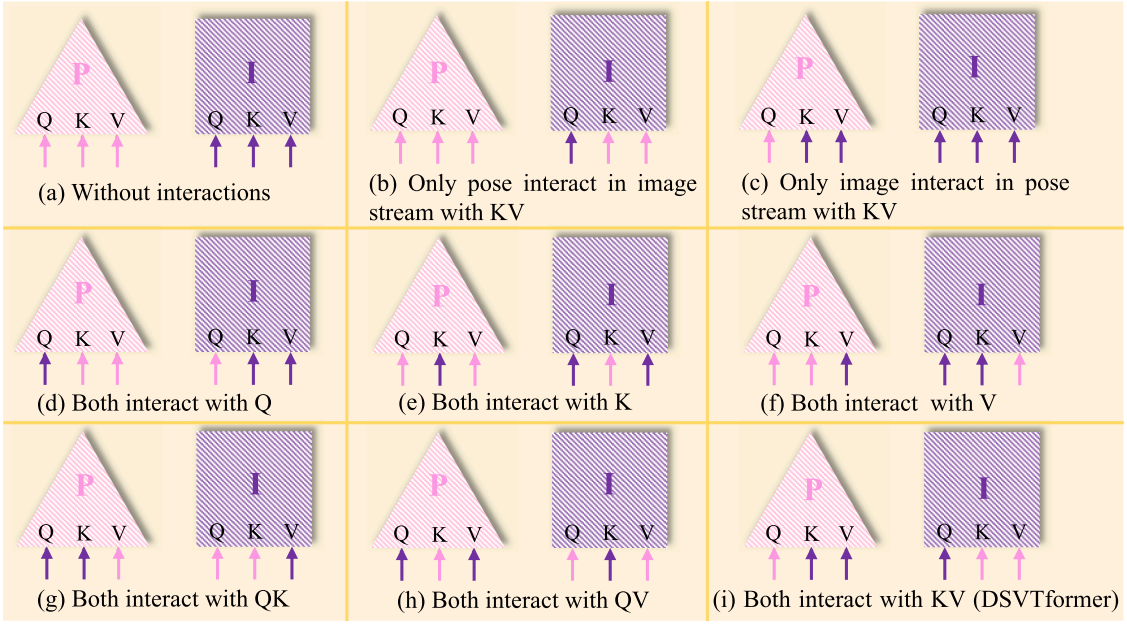


Fig. 6. Variants of dual-stream fusion strategies. Triangles and squares denote pose and image streams with Q/K/V inputs. Pink and purple arrows indicate pose and image features, respectively. DSVTformer adopts the (i) configuration, which achieves the best performance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

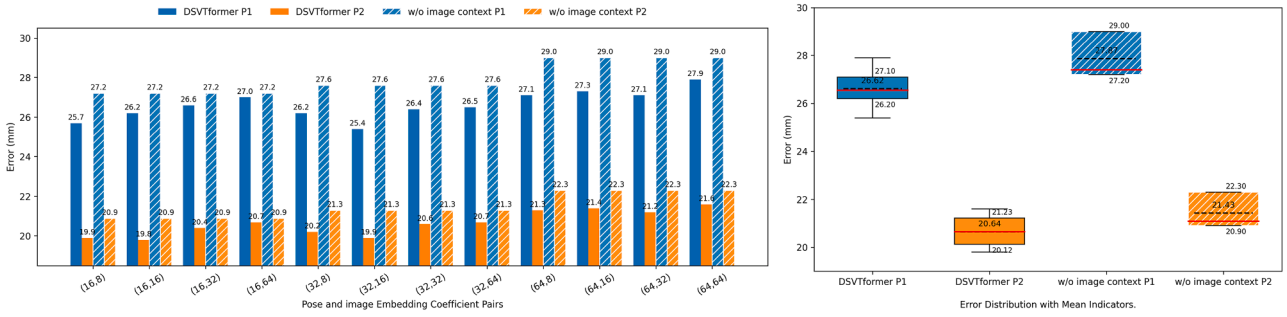


Fig. 7. Ablation study on pose and image embedding coefficients and the corresponding error distributions. Left: We report the errors under Protocols P1 and P2 for different combinations of pose and image embedding coefficients, where each pair on the horizontal axis denotes a specific (pose, image) coefficient setting. Solid bars indicate DSVTformer, while hatched bars correspond to the variant without image context. Right: Error distribution with mean indicators aggregated over all coefficient settings, providing a statistical comparison between DSVTformer and the variant without image context under both P1 and P2. The black dashed lines denote the mean errors, while the red solid lines indicate the median values. Overall, DSVTformer exhibits a more favorable error distribution and achieves competitive or superior performance across a wide range of coefficient settings. In particular, the best performance is obtained when the pose and image embedding coefficients are set to 32 and 16, respectively, indicating that balanced pose–image representations are most effective for cross-modal fusion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

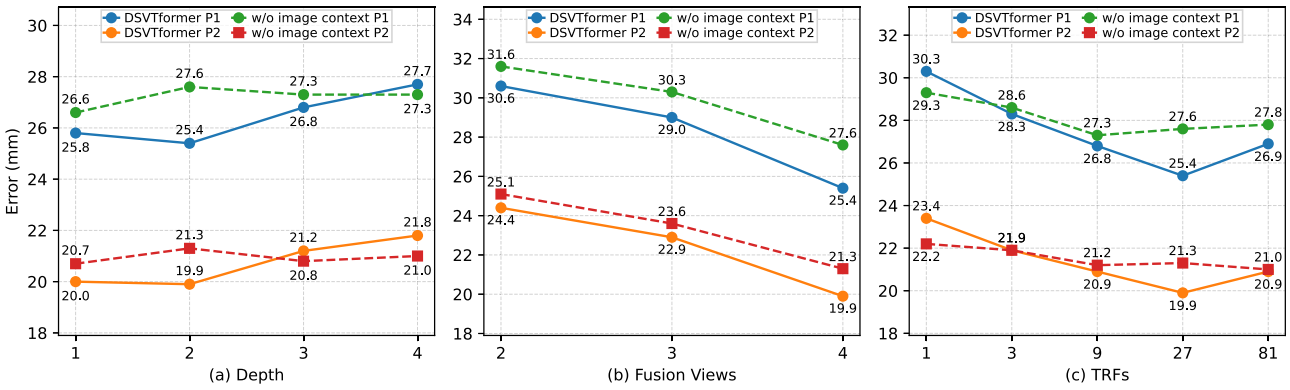


Fig. 8. Ablation study on key architectural hyperparameters: (a) network depth, (b) number of fusion views, and (c) temporal receptive fields (TRFs). MPJPE (*mm*) under Protocols P1 and P2 is reported. Under the default setting of 2 layers, 4 fusion views, and a 27-frame temporal receptive field, DSVTformer achieves the best performance and outperforms its counterpart without image context.

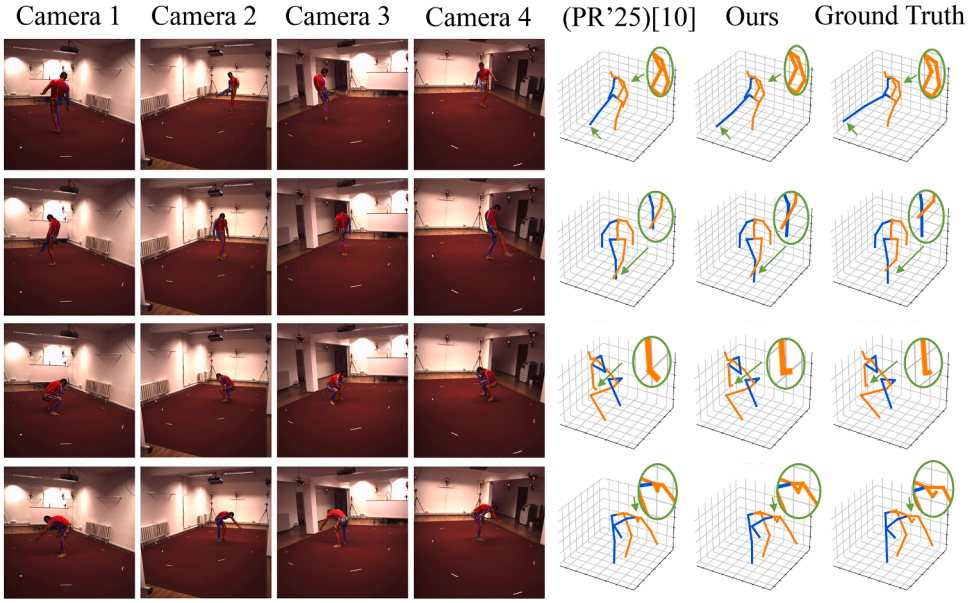


Fig. 9. Visualization results on the Human3.6M dataset. The left part shows four input images captured from different camera views, while the right part presents the corresponding 3D pose estimation results. The enlarged insets emphasize the improved accuracy of skeleton placement and orientation, validating the effectiveness of our dual-stream and sequential spatial-view-temporal attention design for multi-view 3D HPE. Our method effectively captures the complementary information between image features and 2D poses, which is particularly beneficial in challenging scenarios with severe self-occlusion or inaccurate 2D pose detections, leading to more accurate and robust 3D pose predictions.

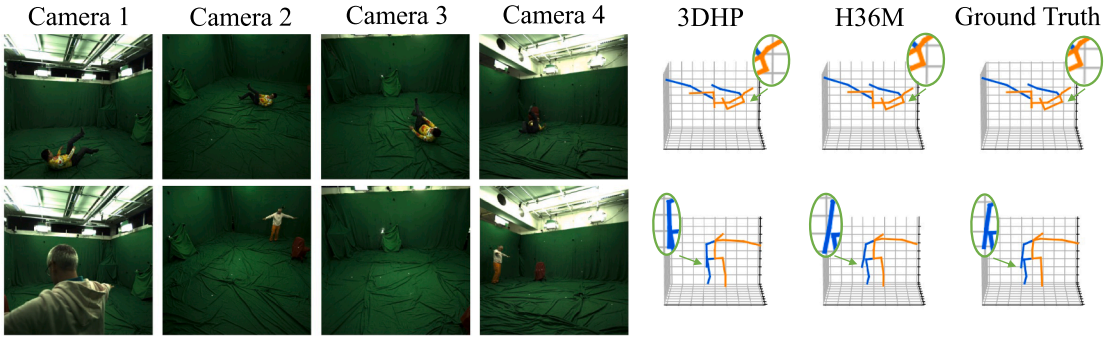


Fig. 10. Visualization results on 3DHP. The dataset mainly comprises green-screen indoor scenes, with some outdoor cases, and includes frequent occlusions and missing body parts. Our model yields accurate 3D poses, with Human3.6M fine-tuning improving performance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

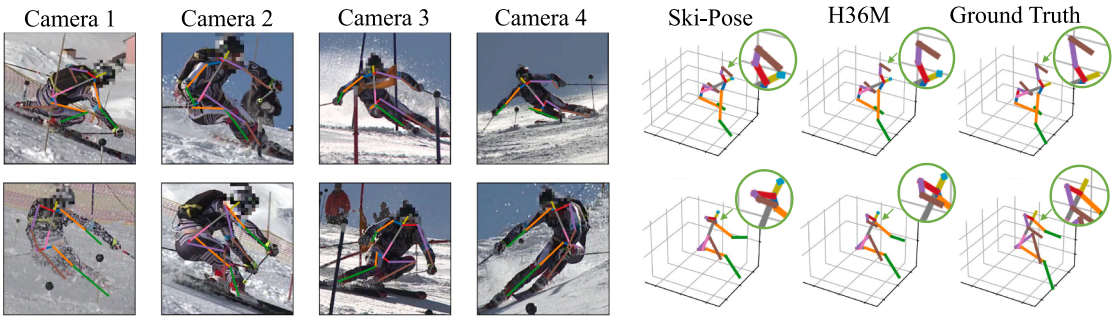


Fig. 11. Visualization results on the Ski-Pose dataset. Our model generates coherent and anatomically plausible 3D poses under challenging outdoor conditions, including motion blur, extreme articulation, and occlusion. Fine-tuning on Human3.6M further enhances performance, demonstrating robustness across training settings.

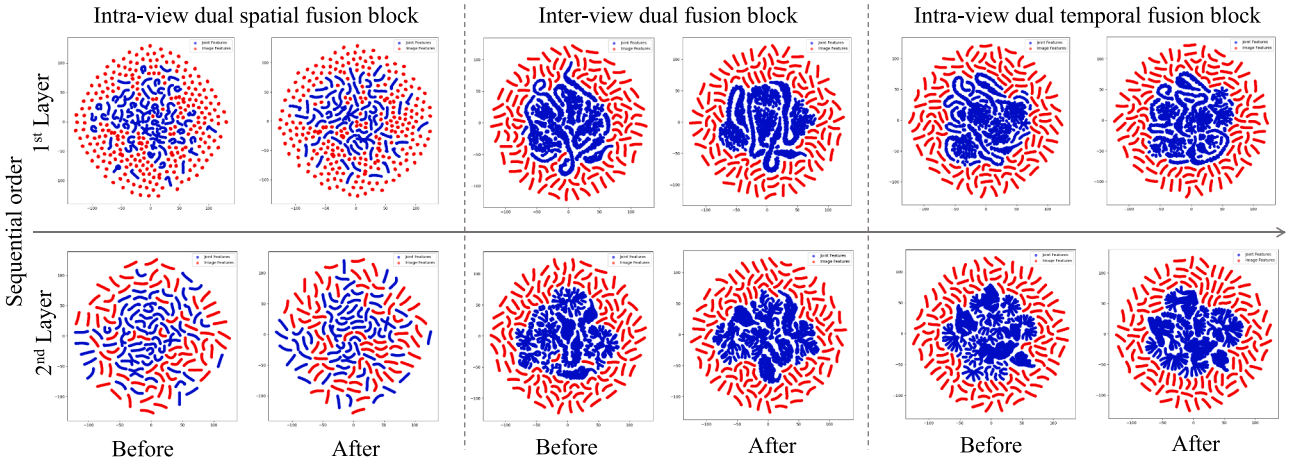


Fig. 12. Dual-modality feature distribution visualization. This shows the t-SNE visualization of the joint (blue) and image (red) feature distributions before and after dual-stream fusion. The figure contains two rows, each corresponding to one encoder layer with three dual-modal fusion blocks: spatial, cross-view, and temporal. In each block, the left plot shows feature distributions before fusion, while the right shows the result after fusion. Across both layers, joint and image features become progressively more compact and aligned, demonstrating the cumulative effect of dual-modality and sequential spatial-view-temporal dual-modal interaction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Ablation study of different dual-stream fusion strategies. Q/K/V denotes the role of image and pose features as inputs in I2PX and P2IX streams, respectively. ✗ indicates no interactions.

ID	Method	Image features in Pose stream (I2PX)	Pose features in Image stream (P2IX)	P1 (*mm*)↓	P2 (*mm*)↓
(a)	Without interactions	✗	✗	35.9	26.0
(b)	Only pose interact in image stream with KV	✗	KV	27.3	21.0
(c)	Only image interact in pose stream with KV	KV	✗	26.1	20.4
(d)	Both interact with Q	Q	Q	35.0	26.6
(e)	Both interact with K	K	K	26.6	21.0
(f)	Both interact with V	V	V	26.3	20.5
(g)	Both interact with QK	QK	QK	34.5	25.6
(h)	Both interact with QV	QV	QV	31.0	24.8
(i)	Both interact with KV (**DSVTformer**)	KV	KV	**25.4**	**19.9**

Table 8
Ablation on 2D pose detectors and stream settings on Human3.6M (P1/P2). The proposed dual-stream formulation exhibits stable performance across different 2D detectors and consistently outperforms the pose-only setting.

Method / 2D Pose Detector	P1 (*mm*)↓	P2 (*mm*)↓
Dual-stream (with image context)		
CPN (Default) [41]	**25.4**	**19.9**
Stacked Hourglass (SH) [53]	26.5	20.4
HRNet [54]	27.0	20.7
Single-stream (pose-only)		
Ours (w/o image context stream)	27.6	21.3

dual-stream DSVTformer consistently outperforms its pose-only counterpart across all evaluated 2D detector settings. This consistent performance gap suggests that the observed improvements stem from the dual-stream design itself, rather than merely from the use of a stronger 2D pose detector, highlighting the effectiveness of integrating complementary visual context. Overall, these results demonstrate that the proposed dual-stream formulation is not only robust to variations in 2D detector quality, but also consistently outperforms single-stream pose lifting baselines by effectively leveraging complementary visual context.

Model configurations. We systematically investigate the influence of key architectural hyperparameters, including pose and image embedding coefficients, network depth, number of fusion views, and temporal receptive fields (TRFs). As shown in Fig. 7, DSVTformer

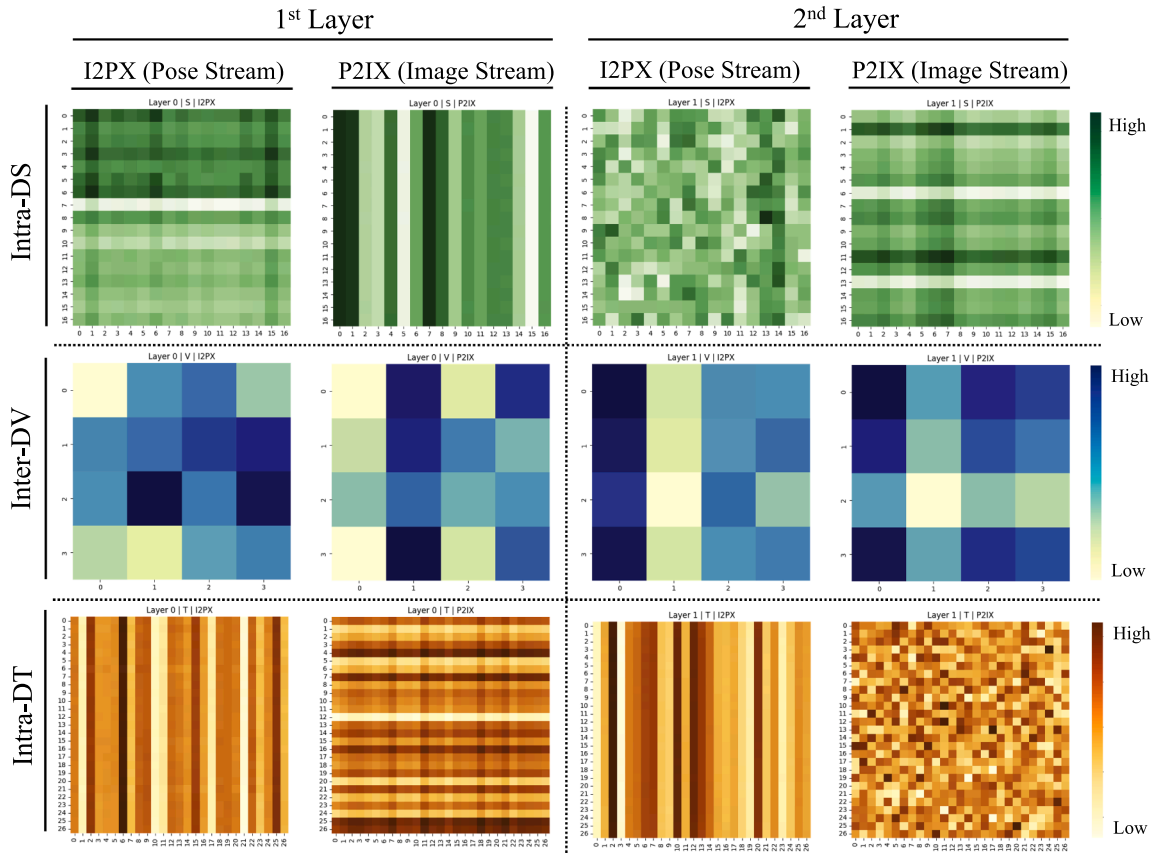


Fig. 13. Attention visualization from I2PX (Image-to-Pose Interaction) and P2IX (Pose-to-Image Interaction) modules at each layer of the dual-stream fusion block. Rows show spatial, cross-view, and temporal correlations; axes denote joints, views, and frames. Columns show attention patterns for each fusion direction. Deeper layers exhibit more structured and informative attention across all dimensions.

demonstrates stable performance trends across a wide range of pose–image embedding coefficient combinations and consistently outperforms the variant without image context under identical settings for both Protocols P1 and P2, indicating low sensitivity to precise coefficient tuning. Beyond mean errors, the right panel of Fig. 7 further shows that DSVTformer yields a more favorable error distribution, with lower means and medians and reduced dispersion compared to the variant without image context, confirming its robustness across different coefficient settings. In particular, balanced pose-image embeddings lead to improved accuracy, with the best performance achieved when the pose and image embedding coefficients are set to 32 and 16, respectively, suggesting that the gains primarily arise from effective cross-modal fusion rather than over-parameterization. Fig. 8 further analyzes the effects of network depth, fusion views, and TRFs: increasing depth initially improves accuracy, with two layers offering the best trade-off; performance steadily benefits from incorporating more views due to richer geometric cues; and extending the TRF reduces errors up to 27 frames, beyond which redundant or noisy temporal information degrades performance. Overall, these results demonstrate that DSVTformer is robust to a broad range of architectural hyperparameters and achieves optimal performance under a balanced configuration with two layers, four fusion views, and a 27-frame temporal receptive field, which is adopted as the default setting in all experiments.

4.5. Visualization results

3D pose visualization. We present qualitative visualizations of our DSVTformer on three datasets. Fig. 9 shows that ours produces 3D pose estimations significantly closer to the ground truth than those by Zhang [10] on Human3.6M. The visual results demonstrate strong robustness to self-occlusion and viewpoint variation. Even when 2D poses contain substantial errors, such as the severely misaligned right leg in Camera 3, the model effectively leverages image context to correct and refine the 3D predictions. This highlights the advantage of our dual-stream fusion design with spatial-view-temporal correlation learning over methods relying solely on 2D poses. Figs. 10 and 11 present qualitative results on 3DHP and Ski-Pose, which pose distinct challenges in controlled indoor and complex outdoor settings, respectively. On both datasets, models fine-tuned from Human3.6M consistently outperform those trained from scratch, indicating that motion priors learned on large-scale indoor data benefit downstream tasks in diverse domains. In both training settings, our method produces more accurate and coherent 3D poses, particularly under severe occlusions, view truncation, and extreme articulations. These results validate the effectiveness of our DSVTformer with a spatial-view-temporal correlation learning framework in handling real-world scenarios by using complementary cues from both images and 2D poses.

Dual-modality feature distribution visualization. To qualitatively assess the dual-modality interactions in our dual-stream fusion blocks, we visualize the distributions of joint and image features using t-SNE [55]. Fig. 12 illustrates two layers of interaction, each comprising spatial, view, and temporal fusion stages. In each stage, feature distributions before and after fusion are shown. Across both layers, features evolve from scattered and modality-separated to increasingly aligned, with spatial fusion promoting structural organization, cross-view fusion enhancing view invariance, and temporal fusion yielding compact, semantically aligned representations. This progressive refinement demonstrates the scalability of our design. Sequential dual-stream cross-modal and intra-modal fusion across layers effectively improves information flow, mitigates modality discrepancy, and produces unified features, en-

hancing spatial accuracy, view consistency, and temporal coherence for 3D pose estimation.

Attention visualization. Fig. 13 presents the attention maps from each layer of dual-stream fusion modules for both I2PX (Image-to-Pose Interaction) and P2IX (Pose-to-Image Interaction). As the network deepens, spatial, cross-view, and temporal attention become more structured and informative, indicating refined correlations through sequential fusion. In shallow layers, I2PX and P2IX focus on local joints, views, and frames, while guiding each other to learn complementary attention patterns. Deeper layers shift toward global representations, capturing spatial structures, cross-view consistency, and temporal dynamics. I2PX highlights fine-grained temporal and structural cues, whereas P2IX captures broader, view-consistent patterns. This complementarity demonstrates the synergy between pose and image streams and underscores the effectiveness of our sequential fusion approach in integrating dual-stream features across space, time, and viewpoint.

5. Conclusion

This work presents DSVTformer, a transformer-based dual-stream framework for multi-view 3D human pose estimation that effectively integrates visual context cues with 2D pose features by modeling their complementary information across spatial, view, and temporal correlation axes. To address the limitations of prior lifting-based methods that overly rely on 2D joint accuracy, we introduce a dual-stream interactive and enhancement mechanism, which enables bidirectional cross-modal interaction and self-modal enhancement, thereby improving feature complementarity between image and pose modalities. On top of this, we propose a sequential, axis-aware decomposition strategy that explicitly captures dual-stream spatial, inter-view, and temporal correlations in a structured manner. This design facilitates fine-grained, progressive reasoning over complex multi-view dynamics, especially under incomplete or unconstrained camera settings.

Empirical results confirm that DSVTformer achieves state-of-the-art accuracy and strong generalization across diverse scenarios. Notably, models trained on Human3.6M can be effectively adapted to other datasets with minimal fine-tuning. Ablations show that incorporating image context is essential for robust 3D pose lifting, and that sequential alternating correlation modeling better exploits cross-modal complementarity. Overall, DSVTformer offers a unified perspective for joint visual-pose modeling, and provides valuable insights for future research in multi-modal representation learning, multi-view reasoning, and 3D human understanding.

While DSVTformer is designed for the clip lengths and view counts commonly used in current multi-view 3D human pose estimation benchmarks, extending the framework to efficiently handle substantially longer temporal contexts remains an open direction. Addressing this challenge may further enhance scalability and broaden the applicability of dual-stream spatial-view-temporal modeling in long-duration and large-scale multi-view scenarios.

CRediT authorship contribution statement

Wanruo Zhang: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Mengyuan Liu:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization; **Wenhao Li:** Writing – review & editing, Software, Conceptualization; **Hong Liu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62373009), the Guangdong S&T Program (Grant No. 2024B0101050002), and the Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (Grant No. RCYX20231211090248064).

References

[1] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3D human pose estimation, in: CVPR, 2018, pp. 7307–7316.
[2] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in: ICCV, 2017, pp. 2602–2611.
[3] Z. Chen, J. Dai, J. Bai, J. Pan, DGFormer: Dynamic graph transformer for 3D human pose estimation, PR 152 (2024) 110446.
[4] V. Davoodnia, S. Ghorbani, M.-A. Carbonneau, A. Messier, A. Etemad, UPose3D: uncertainty-aware 3D human pose estimation with cross-view and temporal cues, in: ECCV, Springer, 2024, pp. 19–38.
[5] J.N. Kundu, S. Seth, P. YM, V. Jampani, A. Chakraborty, R.V. Babu, Uncertainty-aware adaptation for self-supervised 3d human pose estimation, in: CVPR, 2022, pp. 20448–20459.
[6] Q. Cai, X. Hu, S. Hou, L. Yao, Y. Huang, Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser, in: AAAI, vol. 38, 2024, pp. 882–890.
[7] X. Zhang, Q. Cui, Q. Bao, W. Yang, Q. Liao, Geometry-guided diffusion model with masked transformer for robust multi-view 3D human pose estimation, in: ACM MM, 2024, pp. 681–690.
[8] K. Zhou, L. Zhang, F. Lu, X.-D. Zhou, Y. Shi, Efficient hierarchical multi-view fusion transformer for 3D human pose estimation, in: ACM MM, 2023, pp. 7512–7520.
[9] L. Zhang, K. Zhou, F. Lu, X.-D. Zhou, Y. Shi, Deep semantic graph transformer for multi-view 3D human pose estimation, in: AAAI, vol. 38, 2024, pp. 7205–7214.
[10] L. Zhang, K. Zhou, F. Lu, Z. Li, X. Shao, X.-D. Zhou, Y. Shi, ESMformer: error-aware self-supervised transformer for multi-view 3D human pose estimation, PR 158 (2025) 110955.
[11] Q. Zhao, C. Zheng, M. Liu, C. Chen, A single 2D pose with context is worth hundreds for 3d human pose estimation, in: NeurIPS, vol. 36, 2023, pp. 27394–27413.
[12] J. Zhang, M. Liu, H. Liu, G. Wang, W. Li, APP: adaptive pose pooling for 3D human pose estimation from videos, in: ACM MM, 2024, pp. 1672–1681.
[13] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, W. Yang, Exploiting temporal contexts with strided transformer for 3d human pose estimation, IEEE TMM 25 (2022) 1282–1293.
[14] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, Mixste: seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video, in: CVPR, 2022, pp. 13232–13242.
[15] S. Du, Z. Yuan, T. Ikenaga, Kinematics-aware spatial-temporal feature transform for 3D human pose estimation, PR 150 (2024) 110316.
[16] J. Peng, Y. Zhou, P.Y. Mok, KTPFormer: kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation, in: CVPR, 2024, pp. 1123–1132.
[17] M. Kocabas, S. Karagoz, E. Akbas, Self-supervised learning of 3D human pose using multi-view geometry, in: CVPR, 2019, pp. 1077–1086.
[18] E. Remelli, S. Han, S. Honari, P. Fua, R. Wang, Lightweight multi-view 3D pose estimation through camera-disentangled representation, in: CVPR, 2020, pp. 6040–6049.
[19] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, P. Fua, Learning monocular 3D human pose estimation from multi-view images, in: CVPR, 2018, pp. 8437–8446.
[20] M. Burenius, J. Sullivan, S. Carlsson, 3D Pictorial structures for multiple view articulated pose estimation, in: CVPR, 2013, pp. 3618–3625.
[21] G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Harvesting multiple views for marker-less 3d human pose annotations, in: CVPR, 2017, pp. 6988–6997.
[22] H. Qiu, C. Wang, J. Wang, N. Wang, W. Zeng, Cross view fusion for 3D human pose estimation, in: ICCV, 2019, pp. 4342–4351.
[23] Y. He, R. Yan, K. Fragkiadaki, S.-I. Yu, Epipolar transformers, in: CVPR, 2020, pp. 7779–7788.
[24] S. Bultmann, S. Behnke, Real-time multi-view 3D human pose estimation using semantic feedback to smart edge sensors, RSS (2021).
[25] K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov, Learnable triangulation of human pose, in: ICCV, 2019, pp. 7718–7727.
[26] K. Bartol, D. Bojanić, T. Petković, T. Pribanić, Generalizable human pose triangulation, in: CVPR, 2022, pp. 11028–11037.
[27] D.C. Luvizon, D. Picard, H. Tabia, Consensus-based optimization for 3D human pose estimation in camera coordinates, IJCV 130 (3) (2022) 869–882.
[28] B. Gordon, S. Raab, G. Azov, R. Giryes, D. Cohen-Or, FLEX: extrinsic parameters-free multi-view 3D human motion reconstruction, in: ECCV, Springer, 2022, pp. 176–196.

[29] F. Huang, A. Zeng, M. Liu, Q. Lai, Q. Xu, DeepFuse: an IMU-aware network for real-time 3D human pose estimation from multi-view image, in: WACV, 2020, pp. 429–438.
[30] H. Shuai, L. Wu, Q. Liu, Adaptive multi-view and temporal fusing transformer for 3D human pose estimation, IEEE TPAMI 45 (4) (2022) 4122–4135.
[31] S. Mehraban, V. Adeli, B. Taati, MotionAGFormer: enhancing 3D human pose estimation with a transformer-GCNFormer network, in: WACV, 2024, pp. 6920–6930.
[32] Z. Islam, A.B. Hamza, Multi-hop graph transformer network for 3D human pose estimation, JVCIR 101 (2024) 104174.
[33] X. Tong, Z. Zuo, S. Su, J. Wei, X. Sun, P. Wu, Z. Zhao, ST-trans: spatial-temporal transformer for infrared small target detection in sequential images, TGRS 62 (2024) 1–19.
[34] R. Kumar, J. Mendes-Moreira, J. Chandra, Spatio-temporal parallel transformer based model for traffic prediction, TKDD 18 (9) (2024) 1–25.
[35] M. Zhou, S. Han, J. Luo, X. Zhuang, Q. Mao, Z. Li, Transformer-Based and structure-aware dual-stream network for low-light image enhancement, TOMM 21 (10) (2025) 1–24.
[36] E. Yu, J. Lu, K. Zhang, G. Zhang, Enhancing outdoor vision: binocular desnowing with dual-stream temporal transformer, PR 170 (2026) 112075.
[37] B. Tekin, P. Márquez-Neila, M. Salzmann, P. Fua, Learning to fuse 2d and 3d image cues for monocular body pose estimation, in: ICCV, 2017, pp. 3941–3950.
[38] L. Zhao, X. Peng, Y. Tian, M. Kapadia, D.N. Metaxas, Semantic graph convolutional networks for 3d human pose regression, in: CVPR, 2019, pp. 3425–3435.
[39] C. Han, X. Yu, C. Gao, N. Sang, Y. Yang, Single image based 3D human pose estimation via uncertainty learning, PR 132 (2022) 108934.
[40] A. Farhadi, J. Redmon, et al., YOLOv3: an incremental improvement, in: CVPR, vol. 1804, Springer Berlin/Heidelberg, Germany, 2018, pp. 1–6.
[41] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: CVPR, 2018, pp. 7103–7112.
[42] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE TPAMI 36 (7) (2013) 1325–1339.
[43] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3D human pose estimation in the wild using improved CNN supervision, in: 3DV, IEEE, 2017, pp. 506–516.
[44] B. Fasel, J. Spörri, M. Gilgien, G. Boffi, J. Chardonnens, E. Müller, K. Aminian, Three-dimensional body and centre of mass kinematics in alpine ski racing using differential GNSS and inertial sensors, Remote Sens. 8 (8) (2016) 671.
[45] D.P. Kingma, Adam: a method for stochastic optimization,(2014) arXiv:1412.6980.
[46] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, J. Luo, Anatomy-aware 3D human pose estimation with bone-based pose decomposition, IEEE TCSVT 32 (1) (2021) 198–209.
[47] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, Y. Wang, GFPose: learning 3D human pose prior with gradient fields, in: CVPR, 2023, pp. 4800–4810.
[48] W. Li, H. Liu, H. Tang, P. Wang, Multi-hypothesis representation learning for transformer-based 3D human pose estimation, PR 141 (2023) 109631.
[49] F. Zhou, J. Yin, P. Li, Lifting by image–leveraging image cues for accurate 3d human pose estimation, in: AAAI, vol. 38, 2024, pp. 7632–7640.
[50] Q. Zhao, C. Zheng, M. Liu, C. Chen, A single 2D pose with context is worth hundreds for 3D human pose estimation, in: NeurIPS, vol. 36, 2024.
[51] B. Wandt, M. Rudolph, P. Zell, H. Rhodin, B. Rosenhahn, CanonPose: self-supervised monocular 3D human pose estimation in the wild, in: CVPR, 2021, pp. 13294–13304.
[52] M. Gholami, A. Rezaei, H. Rhodin, R. Ward, Z.J. Wang, Self-supervised 3D human pose estimation from video, Neurocomputing 488 (2022) 97–106.
[53] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: ECCV, Springer, 2016, pp. 483–499.
[54] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.
[55] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, JMLR 9 (11) (2008).