USTEP: Spatio-Temporal Predictive Learning Under a Unified View

Cheng Tan ⬤, *Graduate Student Member, IEEE*, Jue Wang, Zhangyang Gao ⬤, Siyuan Li ⬤, *Graduate Student Member, IEEE*, and Stan Z. Li ⬤, *Fellow, IEEE*

Abstract—Spatio-temporal predictive learning plays a crucial role in self-supervised learning, with wide-ranging applications across a diverse range of fields. Previous approaches for temporal modeling fall into two categories: recurrent-based and recurrent-free methods. The former, while meticulously processing frames one by one, neglect short-term spatio-temporal information redundancies, leading to inefficiencies. The latter naively stack frames sequentially, overlooking the inherent temporal dependencies. In this paper, we re-examine the two dominant temporal modeling approaches within the realm of spatio-temporal predictive learning, offering a unified perspective. Building upon this analysis, we introduce USTEP (Unified Spatio-TEmporal Predictive learning), an innovative framework that reconciles the recurrent-based and recurrent-free methods by integrating both micro-temporal and macro-temporal scales. Extensive experiments on a wide range of spatio-temporal predictive learning demonstrate that USTEP achieves significant improvements over existing temporal modeling approaches, thereby establishing it as a robust solution for a wide range of spatio-temporal applications.

Index Terms—Self-supervised learning, spatiotemporal predictive learning, convolutional neural networks, computer vision.

I. INTRODUCTION

IN AN era where data is continually streaming in, there is an increasing demand to not only understand the present but to also predict the future. By leveraging historical video data, spatio-temporal predictive learning strives to forecast subsequent sequences in an unsupervised manner [1], [2], [3], [4], [5], [6], [7], [8], [9]. With real-world applications extending

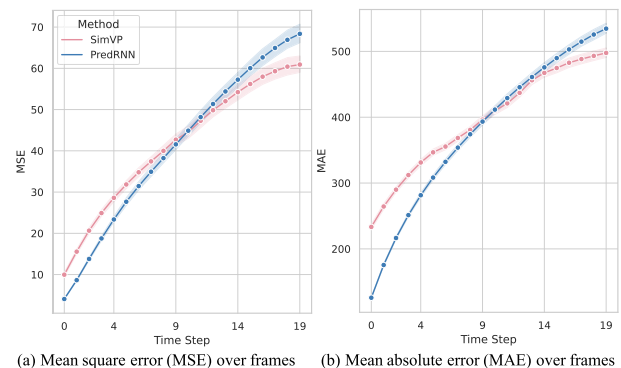(a) Mean square error (MSE) over frames   (b) Mean absolute error (MAE) over frames

Fig. 1. Frame-by-frame MSE/MAE comparison between the representative recurrent-based method PredRNN and the recurrent-free method SimVP on the extended frame task using the KTH dataset. The plot illustrates the differences in performance across individual frames.

from forecasting weather patterns [10], [11], [12], [13], [14] to predicting traffic flows [15], [16] and simulating physical interactions [1], [17], the ramifications of advancements in this promising domain are profound.

The path to achieving accurate spatio-temporal predictions has been fraught with challenges. Traditional approaches have typically oscillated between two primary temporal modeling methodologies: recurrent-based and recurrent-free methods. The recurrent-based methods [10], [16], [18], [19], [20], [21], [22], [23], [24] meticulously process frames one by one, ensuring that temporal relationships across each timestep are captured. Yet, they often grapple with inefficiencies arising from the redundant short-term spatio-temporal information and challenges in preserving global information from preceding time steps. Conversely, the recurrent-free methods [25], [26], [27], [28], while alleviating the inefficiencies of their recurrent counterparts, fall short in capturing the inherent temporal dependencies. By stacking frames in a naive manner, these models may overlook the intricate dance of cause and effect played out over time, risking missing the subtle interplay. As shown in Fig. 1, a frame-by-frame MSE/MAE comparison on the KTH dataset illustrates the strengths and weaknesses of these two approaches. Although the overall average performance metrics for the recurrent-based method PredRNN [18] and the recurrent-free method SimVP [26] are nearly identical (MSE 41.07 versus 41.11, MAE 380.6 versus 397.1), a more granular, frame-by-frame analysis reveals critical insights. PredRNN outperforms SimVP in the initial frames, highlighting its

effectiveness in capturing short-term dependencies. However, in the latter frames, SimVP shows superior performance, underscoring its ability to handle long-term dependencies. The frame-by-frame comparison thus emphasizes the complementary nature of these two approaches. PredRNN's strength in modeling short-term dependencies and SimVP's capability to manage long-term dependencies highlight the inherent trade-offs in choosing one method over the other. This observation underscores the necessity for a unified framework that integrates the advantages of both methodologies, addressing their shortcomings.

In this work, we revisit the foundational principles of temporal modeling in spatio-temporal predictive learning, dissecting the merits and demerits of the prevailing approaches. We introduce the concept of a temporal segment, defined as a subsequence encompassing a series of continuous frames. To refine our understanding further, we formally identify and delineate two temporal scales: the micro-temporal scale, which focuses on immediate, sequential dependencies, and the macro-temporal scale, which encapsulates long-range global patterns. Recurrent-based methods primarily concentrate on micro-temporal scales, adeptly capturing instantaneous interactions but often lacking in long-term insight. Conversely, recurrent-free methods excel in considering macro-temporal scales, effectively capturing broader temporal patterns, but their neglect of immediate temporal dependencies results in a loss of richness in the predicted sequences. This discrepancy between the two paradigms highlights a significant gap in current methodologies. The key challenge is reconciling the redundant per-frame processing of recurrent-based methods with the naive stacking of recurrent-free methods into a unified spatio-temporal learning framework that can effectively model both micro- and macro-temporal scales.

To bridge this gap, we introduce USTEP (Unified Spatio-TEmporal Predictive learning), a novel framework that takes into account both micro- and macro-temporal scales. By doing so, USTEP achieves a balanced trade-off between predictive performance and computational efficiency. The architecture of USTEP is designed to integrate seamlessly the strengths of both recurrent-based and recurrent-free methods, while also introducing new mechanisms that enhance the model's ability to generalize across various spatio-temporal scales. As illustrated in Fig. 2, recurrent-based methods focus solely on the micro-temporal scale, processing one frame at a time; recurrent-free methods consider only the macro-temporal scale, treating the entire sequence as a single unit. USTEP defines micro-temporal segments as several consecutive frames and considers the entire input sequence as a macro-temporal segment, maintaining long-range dependencies as context. This dual-scale approach allows USTEP to learn and update both short-term and long-term temporal relationships dynamically. We conduct a diverse range of spatio-temporal predictive tasks and the experimental results demonstrate its superior performance, not just in terms of accuracy but also in computational efficiency. We find that USTEP achieves state-of-the-art performance with moderate computational resources. Such efficiency and effectiveness establish USTEP

(a) Recurrent-based temporal modeling

(b) Recurrent-free temporal modeling

(c) Unified temporal modeling
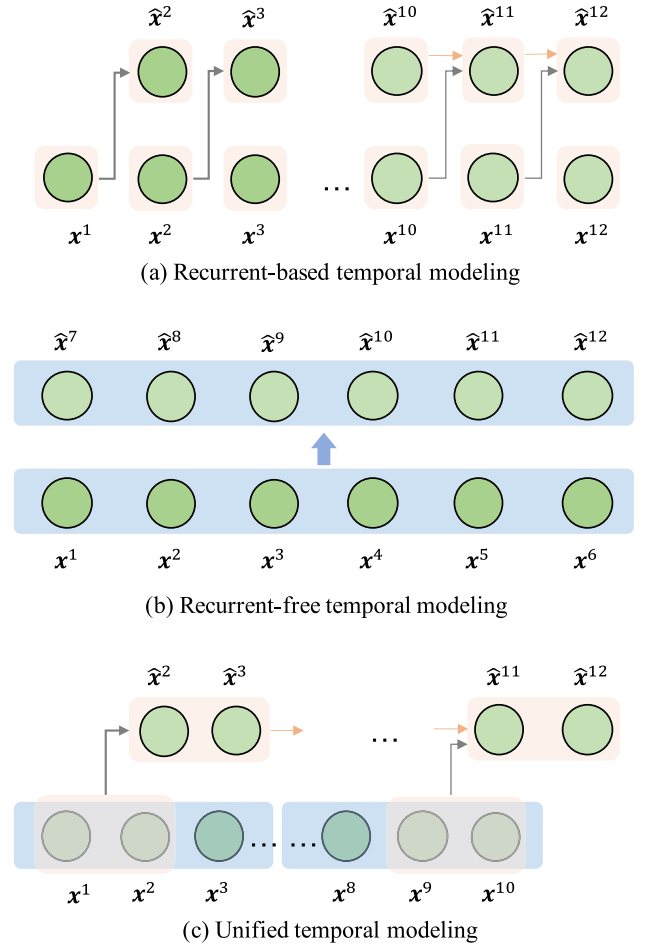
Fig. 2. Temporal modeling comparison between recurrent-based, recurrent-free and our unified temporal modeling.

as an exceptionally powerful solution, particularly suited to addressing the multifaceted challenges of real-world applications.

A preliminary version of this work was published in [25], focusing on advancing temporal attention, and on the exploration of open-source software for spatiotemporal learning [28] (github.com/chengtan9907/OpenSTL). This current journal paper extends them in the following aspects:

1) Leveraging insights gained from the comprehensive spatiotemporal benchmark in OpenSTL [28], this paper provides an in-depth review of prevailing approaches to spatiotemporal predictive learning. A novel perspective is introduced by categorizing these methods according to their operational scales—specifically, micro-temporal and macro-temporal scales. This classification enables a nuanced analysis of their limitations and strengths, fostering a more coherent understanding of the landscape of spatiotemporal predictive learning.

2) Addressing the dichotomy between recurrent-based and recurrent-free learning paradigms, we propose USTEP—a novel, integrated framework designed to harmonize these methodologies by incorporating both micro-temporal and macro-temporal considerations. Using TAU [25] as the basic unit, USTEP represents a significant methodological

advancement, offering a versatile and robust platform for spatiotemporal predictive analysis that transcends the limitations of existing approaches.

3) The efficacy and efficiency of USTEP are rigorously validated across a broad spectrum of spatiotemporal predictive learning tasks. Our experimental investigations encompass tasks with equal frame, extended frame, and reduced frame requirements, demonstrating USTEP's superior performance and adaptability across varying temporal dynamics and spatial configurations.

II. RELATED WORK

A. Recurrent-Based Method

Recurrent-based models have made significant strides in spatio-temporal predictive learning. Drawn inspiration from recurrent neural networks [29], VideoModeling [30] incorporates language modeling techniques and employs quantization of patches into a dictionary for recurrent units. ConvLSTM [10] leverages convolutional neural networks to model the LSTM architecture. PredNet [19] persistently predicts future video frames using deep recurrent convolutional neural networks with bottom-up and top-down connections. PredRNN [18] proposes a Spatio-temporal LSTM (ST-LSTM) unit that extracts and memorizes spatial and temporal representations simultaneously, and its following work PredRNN++ [20] further introduces gradient highway unit and Casual LSTM to capture temporal dependencies adaptively. E3D-LSTM [22] designs eidetic memory transition in recurrent convolutional units. PredRNN-v2 [21] has expanded upon PredRNN by incorporating a memory decoupling loss and a curriculum learning technique. However, these models struggle with capturing long-term dependencies. Moreover, they tend to be computationally intensive, especially when scaled to high-dimensional data, limiting their practical applicability.

B. Recurrent-Free Method

Instead of employing computationally intensive recurrent methods for spatio-temporal predictive learning, alternative approaches such as PredCNN [31] and TrajectoryCNN [32] utilize convolutional neural networks for temporal modeling. SimVP [26], [27] represents a seminal work that incorporates blocks of Inception modules within a UNet architecture. In parallel, the introduction of the TAU [25] represents another leap forward. This innovation underscores the critical role of attention mechanisms in achieving more nuanced and efficient temporal understanding in neural network models. Despite these advances, it is imperative to acknowledge the inherent limitations that accompany these models, particularly in their ability to capture the intricacies of fine-grained temporal dependencies. Additionally, a notable constraint arises from the models' design philosophy concerning output length. Predominantly, these architectures are configured to produce outputs that mirror the length of their inputs, a design choice that, while offering a measure of structural symmetry, introduces rigidity and inherently limits the models' flexibility in decoding.

C. Efficient Recurrent Neural Network

In sequence modeling, RWKV [33] and RetNet [34] revisited the potential of RNNs and propose RNN architectures that can achieve performance comparable to Transformers [35]. Mega [36] proposes a chunk-wise recurrent design, using the moving average equipped gated attention mechanics to capture long-range dependencies in sequential data across various modalities. This design is particularly adept at capturing long-range dependencies across a diverse array of data modalities, highlighting the evolving capabilities of recurrent models in managing complex sequential information. While these prior works demonstrate that well-designed recurrent architectures can be both effective and efficient, USTEP goes a step further by synergizing recurrent and recurrent-free paradigms. This hybrid approach allows USTEP to capture both micro- and macro-temporal scales, offering a nuanced and robust framework for spatio-temporal predictive learning.

III. BACKGROUND

We formally define the spatio-temporal predictive learning problem, inspired by existing works [26], [28]. Consider an observed sequence of frames $\mathcal{X}^{t,T} = \{x^i\}_{t-T+1}^{t}$ at a specific time t, comprising the past T frames. Our objective is to forecast the subsequent T' frames, denoted as $\mathcal{Y}^{t+1,T'} = \{x^i\}_{t+1}^{t+T'}$. Each frame x_i is generally an image in $\mathbb{R}^{C \times H \times W}$, with C being the number of channels, H the height, and W the width. In the tensorial representation, the observed and predicted sequences are represented as $\mathcal{X}^{t,T} \in \mathbb{R}^{T \times C \times H \times W}$ and $\mathcal{Y}^{t+1,T'} \in \mathbb{R}^{T' \times C \times H \times W}$.

Given a model with learnable parameters Θ, we seek to find a mapping $\mathcal{F}_\Theta : \mathcal{X}^{t,T} \mapsto \mathcal{Y}^{t+1,T'}$. This mapping is realized through a neural network model that captures both spatial and temporal dependencies within the data. The model is trained to minimize a loss function $\mathcal{L}$ that quantifies the discrepancy between the predicted and ground-truth future frames, formulated as:

$$\min_{\Theta} \mathcal{L}(\mathcal{F}_\Theta(\mathcal{X}^{t,T}), \mathcal{Y}^{t+1,T'}), \qquad (1)$$

where $\mathcal{L}$ is chosen to evaluate the quality of the predictions in both spatial and temporal dimensions. With the formal problem definition in place, the key challenges in spatio-temporal predictive learning lie in the effective modeling of temporal dependencies, which can be summarized as:

- *Short-term redundancies removal:* Recurrent-based methods process frames one by one, ensuring that temporal relationships across each timestep are captured. However, it often leads to inefficiencies due to redundant processing of short-term spatio-temporal information. These redundancies can cause unnecessary computational overhead and hinder the model's ability to focus on significant temporal changes.
- *Long-term context preservation:* Maintaining global information from preceding time steps is crucial for accurate

long-term predictions. Recurrent-based methods struggle with this as they tend to lose global context. Conversely, recurrent-free methods, which stack frames in a sequence, are more efficient but fail to capture the intricate temporal dependencies necessary for understanding long-term context. This can result in predictions that miss the nuanced interplay of events over time, leading to less accurate forecasts.

To address these challenges, we define a temporal segment as a basic unit comprising a series of continuous frames. We introduce two distinct temporal scales to differentiate between short-term dynamics and long-term context: the micro-temporal scale and the macro-temporal scale. The micro-temporal scale focuses on individual basic temporal segments, capturing fine-grained and immediate interactions. In contrast, the macro-temporal scale encompasses the context with the entire input sequence length, providing a broader context and capturing long-term dependencies. The detailed definitions of them are as follows.

Definition A (Temporal Segment): A temporal segment is defined as a contiguous subsequence of frames extracted from a given spatio-temporal sequence for the purpose of efficient temporal modeling. Formally, let $U_j = \{x^i\}_{t_j}^{t_j + \Delta t - 1}$, where t_j is the starting time of the segment and Δt is the length of the segment measured in time units or number of frames.

By focusing on temporal segments, we aim to capture essential temporal features while mitigating spatio-temporal redundancies that often occur in short-term sequences.

Definition B (Micro-Temporal Scale): The micro-temporal scale refers to the granularity at which a spatio-temporal sequence is partitioned into non-overlapping, contiguous temporal segments for the purpose of efficient and localized temporal modeling. Formally, a sequence $\{x^i\}_{t-T+1}^{t+T'}$ is divided into N micro-temporal segments $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$.

The micro-temporal scale divides the spatio-temporal sequence into non-overlapping temporal segments, emphasizing the independent nature of each segment. This approach captures fine-grained features within each segment while avoiding overlap, thereby ensuring that the detailed information within short-term sequences is accurately preserved and efficiently processed.

Definition C (Macro-Temporal Scale): The macro-temporal scale refers to the granularity at which a spatio-temporal sequence is divided into large temporal segments, each encompassing the length of the entire input sequence. Formally, a sequence $\{x^i\}_{t-T+1}^{t+T'}$ is divided into M macro-temporal segments $\mathcal{V} = \{V_1, V_2, \ldots, V_M\}$, where each segment $V_j = \{U_{j1}, U_{j2}, \ldots, U_{jk}\}$ consists of k micro-temporal segments, and $k\Delta t = \Delta T$ with $\Delta T = T \gg \Delta t$.

The macro-temporal scale aims to capture long-term context by encompassing segments that span the entire length of the input spatio-temporal sequence. The initial macro-temporal segment is equivalent to the recurrent-free approach, providing a broad overview of the input sequence. Subsequent macro-temporal segments advance with a step size equal to that of the micro-temporal segments by updating the long-term context, ensuring a comprehensive and continuous understanding of long-term dependencies across the entire sequence.

A. Recurrent-Based Temporal Modeling

In recurrent-based temporal modeling [10], [18], [37], there exists a predominant focus on the intricacies of micro-temporal scales, to the exclusion of broader, macro-temporal dynamics. Formally, each micro-temporal segment U_i consists solely of a single frame $\{x^i\}$ with $\Delta t = 1$. The modeling approach can be expressed as follows:

$$\widehat{U}_{i+1} = \begin{cases} \mathcal{F}_\Theta(U_i, H_{i-1}), & \text{if } t - T + 1 \leq i \leq t, \\ \mathcal{F}_\Theta(\widehat{U}_i, H_{i-1}), & \text{otherwise,} \end{cases} \quad (2)$$

where H_{i-1} is the hidden state from the preceding frame. The model operates in two distinct phases:

- *Reconstruction Phase:* For historical frames ($t - T + 1 \leq i \leq t$), the model learns to reconstruct the next frame $\widehat{U}_{i+1}$ based on the current ground-truth frame U_i and the hidden state H_{i-1} from the preceding frame.
- *Prediction Phase:* For future frames $i > t$, the model uses the hidden state H_{i-1} and the last predicted frame $\widehat{U}_i$ to predict the next frame $\widehat{U}_{i+1}$.

In both phases, the efficacy of the model for either reconstructing or predicting frames is contingent upon the effective and straightforward learning of the hidden state H_{i-1} from the preceding frame.

B. Recurrent-Free Temporal Modeling

In recurrent-free temporal modeling [25], [26], the focus shifts entirely to the macro-temporal scale, bypassing any micro-temporal segments. Specifically, each macro-temporal segment V is defined as a sequence of T consecutive frames, with $\Delta T = T$. The recurrent-free temporal modeling approach can be mathematically expressed as follows:

$$\widehat{V}_2 = \mathcal{F}_\Theta(V_1), \quad (3)$$

where $V_1 = \{x^i\}_{t-T+1}^t$ is the historical frames, and $V_2 = \{x^i\}_{t+1}^{t+T'}$ is the ground-truth future frames, $\widehat{V}_2$ is the predicted future frames by the model $\mathcal{F}_\Theta$. The model operates in a single phase, where the model learns to predict the future frames $\widehat{V}_2$ based on the historical frames V_1. It is worth noting that here the output frames have the same length as the input frames.

By working with macro-temporal segments, recurrent-free temporal modeling exhibits computational advantages, as it can process multiple frames in parallel. It excels in capturing global patterns over the entire temporal window by taking the macro-temporal segment into account. However, it falters in handling intricate temporal dependencies, primarily because it lacks micro-temporal granularity. The fixed size of macro-temporal segments is inflexible and limits the practical applicability.

C. Summary

In summary, we have dissected the foundational principles underlying temporal modeling in both recurrent-based and recurrent-free methodologies. These approaches are distinguished by their primary operational focus on different temporal scales—namely, the micro-temporal and macro-temporal scales.
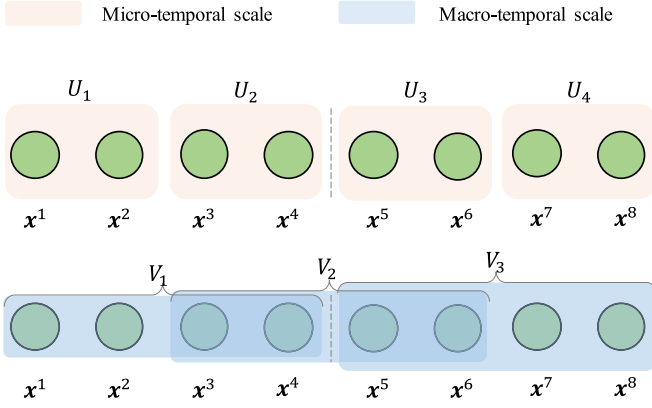
Fig. 3. The illustration of micro- and macro-temporal scales. Here we take a $4 \to 4$ frames prediction as an example. Each green circle represents an individual frame. Micro-temporal scales (*in red*) divide the sequence into non-overlapping temporal segments, containing a few consecutive frames. Macro-temporal scale (*in blue*) further divides the sequence into large temporal segments. The number of $\mathcal{V}$ is $|\mathcal{U}| - 1$.

While recurrent-based temporal modeling focuses on micro-temporal scales, recurrent-free temporal modeling shifts focus to the macro-temporal scale, processing sequences of frames as a whole.

IV. USTEP: UNIFIED SPATIO-TEMPORAL PREDICTIVE LEARNING

To overcome the limitations inherent in both recurrent-based and recurrent-free modeling, we propose USTEP, a unified framework designed to harmoniously integrate micro- and macro-temporal scales, aiming to maximize the effectiveness of spatio-temporal predictive learning.

A. Temporal Scale Sets

The initial stage in USTEP involves dividing the input frame sequence into two separate sets, corresponding to the micro-temporal and macro-temporal scales, which are denoted as $\mathcal{U}$ and $\mathcal{V}$, respectively, as illustrated in 3. In the micro-temporal scale set $\mathcal{U}$, each temporal segment U_i is constructed to contain a few consecutive frames, facilitating the capture of fine-grained spatio-temporal information. The length of each U_i is determined by Δt, which is chosen to balance the trade-off between temporal granularity and computational efficiency. For the macro-temporal scale set $\mathcal{V}$, we employ a sliding window approach to construct larger temporal segments. Each macro-temporal segment V_i contains multiple non-overlapping segments U_i. The sliding window moves in steps of size Δt, ensuring that the macro-temporal segments are constructed in a manner consistent with the micro-temporal scale. As shown in Fig. 4(a), the macro-temporal segments are constructed using both the previous and the current micro-temporal segments, enabling the model to eliminate short-term redundancies while preserving long-term context. This hierarchical approach ensures that the model can effectively capture both immediate and long-term temporal dependencies.

B. Single Segment-Level Temporal Modeling

Upon segmenting the input frame sequence into the micro- and macro-temporal sets, $\mathcal{U}$ and $\mathcal{V}$, the next step is to perform temporal modeling at the single-segment level. This division engages in temporal modeling at an individual segment level, with a concerted effort to achieve a harmonious representation across both temporal dimensions. The core ambition of this phase is to derive hidden states from each temporal segment that are not only compatible across the micro- and macro-temporal scales but are also cohesively aligned within the same feature space that maintains uniform dimensionality in a recurrent-free manner.

To facilitate this objective, temporal segments U_i and V_i, representative of both the micro and macro scales, respectively, are transposed into a shared feature domain, articulated with dimensions $C' \times H \times W$. This transposition is executed through the deployment of two specialized recurrent-free modules, denoted as $F_{\theta_1}^U$ and $F_{\theta_2}^V$. Each module is parameterized with a set of learnable parameters θ_1, θ_2. These modules transform the original segments to corresponding hidden states, denoted as $\overline{U}_i, \overline{V}_i \in \mathbb{R}^{C' \times H \times W}$, according to the following equations:

$$\overline{U}_i = F_\theta^U(U_i), \quad \overline{V}_i = F_\theta^V(V_i), \tag{4}$$

By congruently aligning both micro- and macro-temporal segments within the same feature dimension C', we establish a unified representation space where the hidden states from both temporal scales can be directly integrated. Such a unified feature space enables us to leverage the complementary strengths of micro- and macro-temporal modeling, thus enhancing the overall performance. Analogous to the concept of channel mixing in MetaFormer [38], single segment-level temporal modeling can be viewed as channel mixing at the micro- and macro-temporal scales individually. This approach allows for the independent processing of fine-grained and broad contextual information.

C. Cross Segment-Level Temporal Modeling

Once the unified hidden states are obtained from both micro- and macro-temporal segments, the next challenge is to harmoniously integrate these at the cross-segment level. Our approach is to leverage the advantages of both micro- and macro-temporal scales, capturing fine-grained detail while maintaining a global perspective, to enhance the predictive capability of our model, USTEP.

The macro-temporal scale hidden states, $\overline{V}_i$, are processed using a gating mechanism as follows:

$$\begin{aligned} \boldsymbol{g}_i &= \sigma(W_v * \overline{V}_i + b_v), \\ \boldsymbol{h}_i^V &= \overline{V}_i + \boldsymbol{g}_i \odot \boldsymbol{h}_{i-1}^V, \end{aligned} \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, $*$ and $\odot$ represent the convolution operator and the Hadamard product, respectively. The parameters W_v and b_v are the convolution weights and bias. The context gate, $\boldsymbol{g}_i$, controls the flow of historical macro-temporal scale information and preserves long-term context by regulating the integration of new information with the previously accumulated macro-temporal scale hidden states.

(a) Temporal segment partition
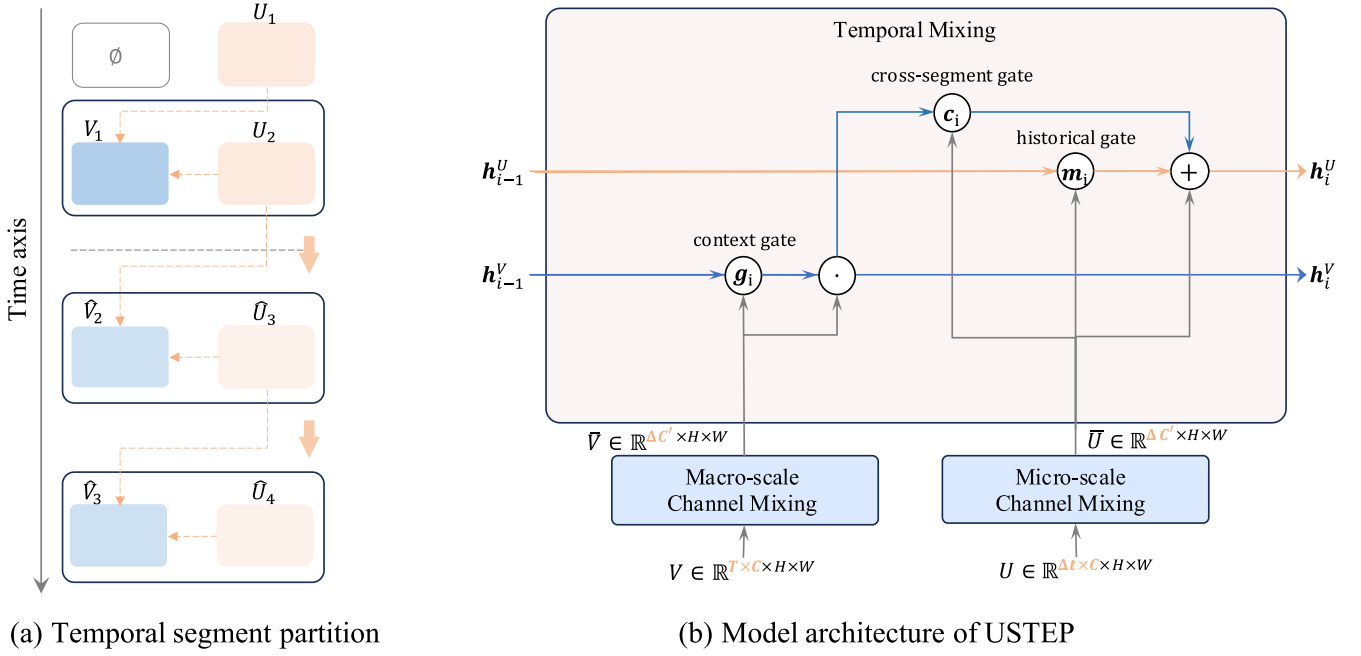
(b) Model architecture of USTEP

Fig. 4. Illustration of USTEP's unified spatio-temporal predictive learning framework. (a) Temporal segment partition: USTEP constructs macro-temporal segments using both the previous and the current micro-temporal segments. This approach allows the model to eliminate short-term redundancies while preserving long-term context. (b) Detailed architecture of USTEP: The framework consists of two specialized recurrent-free modules, $F_{\theta_1}^U$ and $F_{\theta_2}^V$, which handle channel mixing for micro- and macro-temporal scales, respectively. Hidden states from both scales are integrated through a gating mechanism and cross-segment-level temporal modeling, ensuring comprehensive spatio-temporal predictive learning.

Subsequently, the micro-temporal scale and the processed macro-temporal hidden states are then integrated:

$$\boldsymbol{m}_i = \sigma(W_u * \overline{U}_i + b_u),$$
$$\boldsymbol{c}_i = \sigma(W_c * \overline{U}_i + b_c),$$
$$\boldsymbol{h}_i^U = \overline{U}_i + \boldsymbol{m}_i \odot \boldsymbol{h}_{i-1}^U + \boldsymbol{c}_i \odot \boldsymbol{h}_i^V, \quad (6)$$

where W_u, W_c, b_u, and b_c are the convolution weights and biases. The historical gate $\boldsymbol{m}_i$ controls the integration of the historical micro-temporal information, ensuring that the model effectively utilizes past micro-temporal hidden states. The cross-segment gate $\boldsymbol{c}_i$ manages the incorporation of the macro-temporal hidden state, ensuring that long-term context is integrated with the immediate temporal features. This dual-gate mechanism ensures that the model leverages both short-term and long-term dependencies, enhancing its predictive capabilities and overall performance.

The detailed schematic illustration of the temporal modeling learning process within USTEP is depicted in Fig. 4(b). For every single segment, regardless of whether it belongs to the micro-temporal or macro-temporal scale, a recurrent-free approach is employed to swiftly capture coarse temporal dependencies. In contrast, when considering the relationships across segments, a recurrent-based approach is sequentially applied to discern finer temporal dependencies. Notably, during this stage, temporal dependencies from different scales are harmoniously fused, ensuring a comprehensive temporal understanding. This hierarchical and integrative approach allows USTEP to achieve a delicate balance between capturing immediate temporal nuances and understanding broader temporal patterns.

In the training phase, our proposed USTEP method adopts a partitioning approach to divide the input sequences into x and y, which aligns with the training strategy typically employed by recurrent-based methods. However, USTEP offers a flexible approach compared to traditional methods, which often use a fixed time step of $\Delta t = 1$. The flexibility in USTEP allows for different time step sizes, accommodating various temporal resolutions and capturing temporal dependencies at different granularities. During the inference phase, the model engages in iterative operations for a predetermined number of steps. These steps encompass both those required for observing sequences x and the actual steps required for generating the predictions.

The algorithmic structure for the training process of our proposed USTEP is presented in Algorithm 1, providing a comprehensive roadmap for its implementation. Notably, the partition function supports to handle with videos with flexible lengths by adding up a pad.

During the inference phase, USTEP engages in a series of iterative operations spanning a pre-defined number of steps, which include both the observation of sequence x and the execution of steps necessary for generating forward-looking predictions. This iterative approach ensures a dynamic and responsive prediction process, tailored to the temporal characteristics of the input data.

The essence of this inference process is captured within the pseudocode provided in Algorithm 2 below. This algorithm is not just a sequence of steps but a manifestation of the USTEP framework's core capabilities, illustrating its approach to integrating

Algorithm 1. Pseudocode of Training.

```
def train(data, delta_t, delta_T):
    x = data[:, :-delta_t]
    y = data[:, delta_t:]
    # partition: micro and macro sets
    u, v = partition(x, delta_t, delta_T)

    # single segment-level: recurrent-free
    u, v = f_u(u), f_v(v)
    # cross segment-level: recurrent-based
    pred = []
    h_v, h_u = 0, 0
    for i in range(len(u)-1):
        h_v = macro_func(v[i], h_v)
        h_u = micro_func(u[i+1], h_u, h_v)
        pred.append(h_u)

    loss = loss_func(pred, y)
    return loss
```

Algorithm 2. Pseudocode of Inference.

```
def inference(x, delta_t, delta_T, n_step):
    # x: B x T x C x H x W
    # partition: micro and macro sets
    u, v = partition(x, delta_t, delta_T)

    # single segment-level: recurrent-free
    u, v = f_u(u), f_v(v)
    # cross segment-level: recurrent-based
    pred = []
    h_v, h_u = 0, 0
    for i in range(n_step-1):
        h_v = macro_func(v[i], h_v)
        h_u = micro_func(u[i+1], h_u, h_v)
        # practical prediction
        if i >= len(u):
            pred.append(h_u)

    return pred
```

and operationalizing the micro and macro temporal scales in a seamless and efficient manner.

V. EXPERIMENTS

We evaluate the efficacy of USTEP across three prevalent types of spatiotemporal prediction tasks:

- *Equal Frame Task:* The number of output frames matches that of the input frames. This task type inherently favors recurrent-free temporal modeling approaches due to its structured nature.
- *Extended Frame Task:* The count of output frames substantially surpasses that of the input frames. This type of task is generally more compatible with recurrent-based temporal modeling approaches, allowing for more flexible, frame-by-frame predictions.
- *Reduced Frame Task:* Diverging from the former, this task necessitates fewer output frames than the input frames. By mitigating the impact of cumulative errors, this task directly evaluates the model's capability in learning historical frames.

Datasets: To rigorously assess the performance and applicability, we undertake a comprehensive quantitative evaluation across a diverse range of datasets, meticulously curated to encompass both synthetic and real-world scenarios:

- *Moving MNIST* [39] is a synthetic dataset consisting of two digits moving within the 64×64 grid and bouncing off the boundary. It is a standard benchmark in spatiotemporal predictive learning.
- *Human 3.6 M* [40] is a 3D human motion capture dataset for fitness, close human interactions, and self-contact. This dataset contains 3.6 million human poses and corresponding images, 11 professional actors (6 male, 5 female), and 17 scenarios (discussion, smoking, taking photos, talking on the phone, etc.).
- *Weather Benchmark* [12] This dataset contains various types of climatic data from 1979 to 2018. The raw data is regrind to low resolutions, we here choose $5.625°$ (32×64 grid points) resolution for our data. Since the complete data is very large and includes massive climatic attributes like geopotential, temperature, and other variables, we

specifically chose the global temperature prediction task in evaluation.

- *Caltech Pedestrian* is a driving dataset focusing on detecting pedestrians. It consists of approximately 10 hours of 640×480 videos taken from vehicles driving through regular traffic in an urban environment. We follow the protocol of PredNet [19] and CrevNet [41] for pre-processing, training, and evaluation.
- *SEVIR* [42] is a comprehensive storm event imagery dataset comprising over 10,000 weather events. Each event is represented by image sequences covering a spatial extent of 384 km $\times$ 384 km and spanning a temporal duration of 4 hours.
- *UCF Sports* [43] is a dataset featuring a collection of human actions performed in various sports scenarios. These actions are typically recorded from footage broadcast on television channels.
- *KTH* [44] contains 25 individuals performing six types of actions. Following [22], [45], we use persons 1-16 for training and 17-25 for testing. Models are trained to predict the next 20 frames from the previous 10 observations.

We summarize the statistics of the above datasets in Table I, including the number of training samples N_{train} and the number of testing samples N_{test}.

Baselines: We choose the following baselines for comparison: (i) Recurrent-based methods including ConvLSTM [10], SV2P [46], SAVP [47], PredRNN [18], PredRNN++ [20], MIM [16], E3D-LSTM [22], and PredRNNv2 [21]; (ii) Recurrent-free methods including SimVP [26], TAU [25], Uniformer [48], MLP-Mixer [49], ConvNeXt [50] and IAM4VP [14]. For USTEP, we utilize TAU, SimVP, and Uniformer as the single segment-level temporal modeling modules $F_{\theta_1}^{U}$ and $F_{\theta_2}^{V}$ to demonstrate the robustness and versatility of our method. Notably, in USTEP, these modules retain the same architecture as their original models but are configured with fewer channels to ensure a fair comparison under similar computational complexity.

Measurement: We employ Mean Squared Error (MSE), Mean Absolute Error (MAE), Structure Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR) to evaluate the quality of predictions. MSE and MAE estimate the absolute pixel-wise errors, SSIM measures the similarity of structural

TABLE I
THE STATISTICS OF DATASETS

	N_{train}	N_{test}	(C, H, W)	T	T'
MMNIST	10,000	10,000	(1, 64, 64)	10	10
Human 3.6M	73,404	8,582	(3, 256, 256)	4	4
WeatherBench	2,167	706	(1, 32, 64)	12	12
Kitti&Caltech	3,160	3,095	(3, 128, 160)	10	1
SEVIR	35,718	12,159	(1, 384, 384)	13	12
UCF Sports	6,288	752	(3, 480, 720)	4	1
KTH	4,940	3,030	(1, 128, 128)	10	20/40

The training or testing set has N_{train} or N_{test} samples, composed by T or T' images with the shape (C,H,W).

information within the spatial neighborhoods, and PSNR is an expression for the ratio between the maximum possible power of a signal and the power of distorted noise. LPIPS [51] is a perceptual similarity metric that computes the distance between two images' feature representations in a pre-trained deep network. FVD score [52] is also included to evaluate the performance on the KTH dataset.

Implementation details: We implement the proposed method with the Pytorch framework and conduct experiments on a single NVIDIA-V100 GPU. The AdamW optimizer is utilized with a learning rate of 0.01 and a weight decay of 0.05. The learning rate is chosen from a set of values, $\{1e^{-2}, 5e^{-3}, 1e^{-3}, 5e^{-4}, 1e^{-4}\}$, and the best result for each experiment is reported. It is worth noting that the experimental results in TAU [25] replicate the experimental settings from the previous baselines, which are not fully rigorous. For consistency and comparability, we adhere to the same settings as OpenSTL [28].

A. Equal Frame Task

Under the experimental setup of the Equal frame task paradigm, we meticulously conducted a comprehensive evaluation of the model's performance on three distinct datasets. These datasets, namely Moving MNIST [39], Human3.6M [40], and WeatherBench [12], were judiciously chosen for their inherent property of symmetrical frame sequences, where the number of output frames precisely mirrors that of the input frames. This characteristic facilitates a rigorous assessment of the model's capacity to generate temporally coherent and contextually accurate predictions over varying domains and complexity levels.

The Moving MNIST dataset requires the model to forecast 10 subsequent frames based on an equivalent number of preceding frames. This task challenges the model to understand and predict the dynamics of two moving digits within a frame, encapsulating the complexity of motion patterns and interactions. Similarly, the Human3.6 M dataset, derived from a rich repository of human activities, mandates the prediction of 4 future frames from 4 given frames. This dataset serves as a crucible for evaluating the model's proficiency in capturing and forecasting human motion dynamics. The WeatherBench dataset introduces a unique challenge: predicting 12 future temperature patterns from 12 preceding snapshots. This scenario tests the model's ability to grasp and anticipate complex meteorological evolutions over time. The detailed results are summarized in Table II.

In the Moving MNIST dataset, USTEP w/TAU outperforms the previously top-ranked recurrent-based model, PredRNN++, not only in predictive accuracy but also in computational efficiency. Remarkably, USTEP achieves these superior results while requiring only half the number of parameters and a mere tenth of the FLOPs compared to PredRNN++. In a direct comparison with the leading recurrent-free model, TAU, USTEP not only operates within a comparable range of FLOPs but does so with approximately 42% of TAU's parameters, thereby demonstrating significant improvements across various performance metrics. This leap in efficiency and effectiveness is illustrated through a compelling visualization example presented in Fig. 5, highlighting USTEP's prowess in generating highly accurate predictions in the Moving MNIST challenge.

On the Human3.6 M, a complex dataset capturing a wide array of human activities, USTEP w/TAU continues to demonstrate superior performance. Notably, it employs one-third of the parameters required by its closest competitor, Uniformer, and operates with approximately 12% fewer FLOPs. Moreover, USTEP w/Uniformer stands out as the leading model in all evaluated metrics with about 30% fewer parameters. This significant reduction in computational resources, without a compromise in performance, underscores USTEP's capability to predict human motion with remarkable efficiency.

The assessment on WeatherBench, a challenging dataset for meteorological prediction, further solidifies USTEP's position as a versatile and potent model in the arena of spatiotemporal forecasting. USTEP w/TAU obtains impressive advancements over existing models across a spectrum of metrics. It maintains FLOPs close to the premier recurrent-free model, TAU, while its parameter count is merely 29% of TAU's. Furthermore, when juxtaposed with the leading recurrent-based model, PredRNN, USTEP's parameter efficiency becomes even more pronounced, with its parameters constituting a mere 15% of those required by PredRNN. Among these datasets, USTEP demonstrates significant improvements over TAU, SimVP, and Uniformer in performance metrics. Notably, it achieves these enhancements with a lower parameter count and similar FLOPs, showcasing its robustness and versatility across diverse domains. Through these rigorous evaluations, USTEP showcases its prowess in precise and efficient prediction across various datasets under the equal frame task setting.

B. Extended Frame Task

In the exploration of spatiotemporal predictive models' capabilities, particularly in the context of forecasting a greater number of frames than those observed, the KTH dataset emerges as a pivotal benchmark. The intricacies of this task lie in its demand for models to not only understand and internalize short-term sequences but also to extrapolate these observations into longer, future sequences, thereby amplifying the challenge of accurate prediction.

Upon a detailed examination, as presented in Table III, a distinct pattern emerges. Recurrent-free models, though commendable for their efficiency in terms of both parameters (Params) and computational complexity (FLOPs), tend to fall short in

TABLE II
THE QUANTITATIVE RESULTS OF DIFFERENT APPROACHES ON THE EQUAL FRAME TASK

Dataset	Metric	Method				
Moving MNIST		ConvLSTM	PredRNN	PredRNN++	MIM	E3DLSTM
	MSE	29.80	23.97	22.06	22.55	35.97
	MAE	90.64	72.82	69.58	69.97	78.28
	SSIM ($\times 10^{-2}$)	92.88	94.62	95.09	94.98	93.20
	PSNR	22.10	23.28	23.65	23.56	21.11
	Params	**15.0**	23.8	38.6	38.0	51.0
	FLOPs	56.8	116.0	171.7	179.2	298.9
		PredRNNv2	MLP-Mixer	Uniformer	SimVP	TAU
	MSE	24.13	29.52	30.38	32.15	24.60
	MAE	73.73	83.36	85.87	89.05	71.93
	SSIM ($\times 10^{-2}$)	94.53	93.38	93.08	92.68	94.54
	PSNR	23.21	22.13	22.78	21.84	23.19
	Params	23.9	44.8	46.8	58.0	44.7
	FLOPs	116.6	16.5	16.5	19.4	16.0
		ConvNeXt	IAM4VP	USTEP w/Uniformer	USTEP w/SimVP	USTEP w/TAU
	MSE	26.94	27.04	28.51$_{(-1.87)}$	30.04$_{(-2.11)}$	**21.84**$_{(-2.76)}$
	MAE	77.23	79.70	82.43$_{(-3.44)}$	85.92$_{(-3.13)}$	**63.21**$_{(-8.72)}$
	SSIM ($\times 10^{-2}$)	93.97	93.95	94.32$_{(+1.24)}$	93.59$_{(+0.91)}$	**95.38**$_{(+0.84)}$
	PSNR	22.22	22.19	23.28$_{(+0.50)}$	22.11$_{(+0.27)}$	**24.06**$_{(+0.87)}$
	Params	37.3	35.4	17.8$_{(-61.97\%)}$	24.4$_{(-57.93\%)}$	18.9$_{(-57.72\%)}$
	FLOPs	**14.1**	14.2	16.8$_{(+1.82\%)}$	21.9$_{(+12.89\%)}$	17.7$_{(+10.63\%)}$
Human3.6M		ConvLSTM	PredRNN	PredRNN++	MIM	E3DLSTM
	MSE	125.5	113.2	110.0	112.1	143.3
	MAE	1566.7	1458.3	1452.2	1467.1	1442.5
	SSIM ($\times 10^{-2}$)	98.13	98.31	98.32	98.29	98.03
	PSNR	33.40	33.94	34.02	33.97	32.52
	Params	15.5	24.6	39.3	47.6	60.9
	FLOPs	347.0	704.0	1033.0	1051.0	542.0
		PredRNNv2	MLP-Mixer	Uniformer	SimVP	TAU
	MSE	114.9	116.3	108.4	115.8	113.3
	MAE	1484.7	1497.7	1441.0	1511.5	1390.7
	SSIM ($\times 10^{-2}$)	98.27	98.24	98.34	98.22	98.39
	PSNR	33.84	33.76	34.08	33.73	34.03
	Params	24.6	27.7	11.3	41.2	37.6
	FLOPs	708.0	211.0	74.6	197.0	182.0
		ConvNeXt	IAM4VP	USTEP w/Uniformer	USTEP w/SimVP	USTEP w/TAU
	MSE	113.4	114.7	**106.1**$_{(-2.3)}$	109.8$_{(-6.0)}$	109.5$_{(-3.8)}$
	MAE	1469.7	1489.9	1415.3$_{(-25.7)}$	1472.7$_{(-38.8)}$	**1380.5**$_{(-10.2)}$
	SSIM ($\times 10^{-2}$)	98.28	98.26	98.42$_{(+0.08)}$	98.38$_{(+0.16)}$	**98.45**$_{(+0.06)}$
	PSNR	33.86	33.74	34.20$_{(+0.12)}$	34.11$_{(+0.38)}$	**34.35**$_{(+0.32)}$
	Params	31.4	102.0	**3.4**$_{(-69.91\%)}$	4.5$_{(-89.08\%)}$	3.7$_{(-90.16\%)}$
	FLOPs	157.0	978.0	76.6$_{(+2.68\%)}$	**63.2**$_{(-67.92\%)}$	66.2$_{(-63.63\%)}$
WeatherBench		ConvLSTM	PredRNN	PredRNN++	MIM	E3DLSTM
	MSE	1.521	1.331	1.634	1.784	1.592
	MAE ($\times 10^{-2}$)	79.49	72.46	78.83	87.16	80.59
	RMSE	1.233	1.154	1.278	1.336	1.262
	Params	15.0	23.6	38.3	37.8	51.1
	FLOPs	136.0	278.0	413.0	109.0	169.0
		PredRNNv2	MLP-Mixer	Uniformer	SimVP	TAU
	MSE	1.545	1.255	1.204	1.238	1.162
	MAE ($\times 10^{-2}$)	79.86	70.11	68.85	70.37	67.07
	RMSE	1.243	1.119	1.097	1.113	1.078
	Params	23.6	11.1	12.0	14.7	12.2
	FLOPs	279.0	5.9	7.5	8.0	6.7
		ConvNeXt	IAM4VP	USTEP w/Uniformer	USTEP w/SimVP	USTEP w/TAU
	MSE	1.277	1.988	1.190$_{(-0.014)}$	1.215$_{(-0.023)}$	**1.150**$_{(-0.012)}$
	MAE ($\times 10^{-2}$)	72.20	96.43	67.32$_{(-1.53)}$	69.11$_{(-1.26)}$	**65.83**$_{(-1.24)}$
	RMSE ($\times 10^{-2}$)	1.130	1.410	1.087$_{(-0.010)}$	1.102$_{(-0.011)}$	**1.072**$_{(-0.006)}$
	Params	10.1	12.9	3.3$_{(-72.50\%)}$	4.4$_{(-70.07\%)}$	3.6$_{(-70.49\%)}$
	FLOPs	5.7	**2.4**	7.7$_{(+2.67\%)}$	9.9$_{(+23.75\%)}$	8.2$_{(+22.39\%)}$

The units for Params and FLOPs are M and G. The results of the USTEP series that underperform the default are in blue while outperforming those are in red.
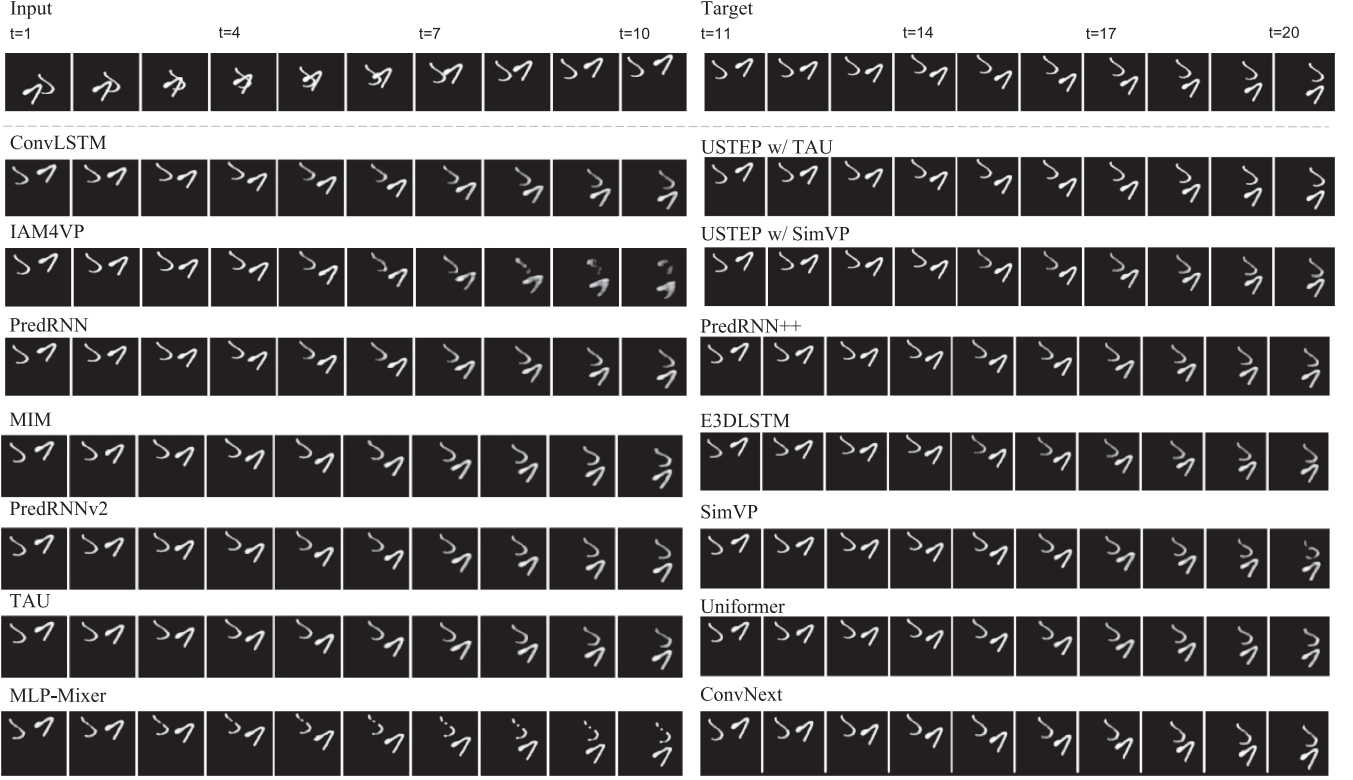
Fig. 5. The qualitative visualization on Moving MNIST.

TABLE III
THE QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE KTH DATASET (10 → 20 FRAMES)

Method	Param	FLOPs	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
ConvLSTM	14.9	1368.0	47.65	445.5	0.8977	26.99	26.69	640.78
PredRNN	23.6	2800.0	41.07	380.6	0.9097	27.95	21.89	436.26
PredRNN++	38.3	4162.0	39.84	370.4	0.9124	28.13	19.87	394.63
MIM	39.8	1099.0	40.73	380.8	0.9025	27.78	**18.81**	1105.97
E3DLSTM	53.5	217.0	136.40	892.7	0.8153	21.78	48.36	1998.82
PredRNNv2	23.6	2815.0	39.57	368.8	0.9099	28.01	21.48	500.30
SV2P	8.3	468.0	64.88	687.2	0.8524	27.77	27.60	433.19
SAVP	17.6	641.0	124.53	867.8	0.8192	26.00	23.41	685.70
MLP-Mixer	20.3	66.6	57.74	517.4	0.8886	25.72	28.80	761.58
ConvNeXt	12.5	63.9	45.48	428.3	0.9037	26.96	26.25	569.33
Uniformer	**11.8**	78.3	44.71	404.6	0.9058	27.16	24.17	543.40
SimVP	12.2	**62.8**	41.11	397.1	0.9065	27.46	26.50	567.72
TAU	15.0	73.8	45.32	421.7	0.9086	27.10	22.86	530.62
USTEP w/Uniformer	12.0$_{(+1.69\%)}$	101.0$_{(+28.99\%)}$	40.23$_{(-4.48)}$	377.5$_{(-27.1)}$	0.9140$_{(+0.0082)}$	28.55$_{(+1.39)}$	20.15$_{(-4.02)}$	420.30$_{(-123.10)}$
USTEP w/SimVP	16.4$_{(+34.43\%)}$	137.0$_{(+118.15\%)}$	39.75$_{(-1.36)}$	372.4$_{(-24.7)}$	0.9135$_{(+0.0070)}$	28.47$_{(+1.01)}$	20.35$_{(-6.15)}$	410.80$_{(-156.92)}$
USTEP w/TAU	12.8$_{(-14.67\%)}$	107.0$_{(+44.99\%)}$	**39.55**$_{(-5.77)}$	**364.9**$_{(-56.8)}$	**0.9165**$_{(+0.0079)}$	**28.98**$_{(+1.88)}$	19.96$_{(-2.90)}$	**393.13**$_{(-137.49)}$

The results of the USTEP series that underperform the default are in blue while outperforming those are in red.

achieving the high-performance benchmarks set by their recurrent-based counterparts. However, within this competitive landscape, USTEP distinctly outperforms the previously established frontrunner, PredRNN++, both in terms of predictive accuracy and computational efficiency. USTEP not only aligns itself with the efficiency metrics typically seen in recurrent-free models but also surpasses them in performance. This juxtaposition is illustrated in Fig. 6, where USTEP w/TAU predicts notably more realistic and sharply defined compared to those of other methods. While TAU, a recurrent-free model, exhibits commendable performance, it struggles to replicate the nuanced details that USTEP w/TAU manages to capture. By integrating
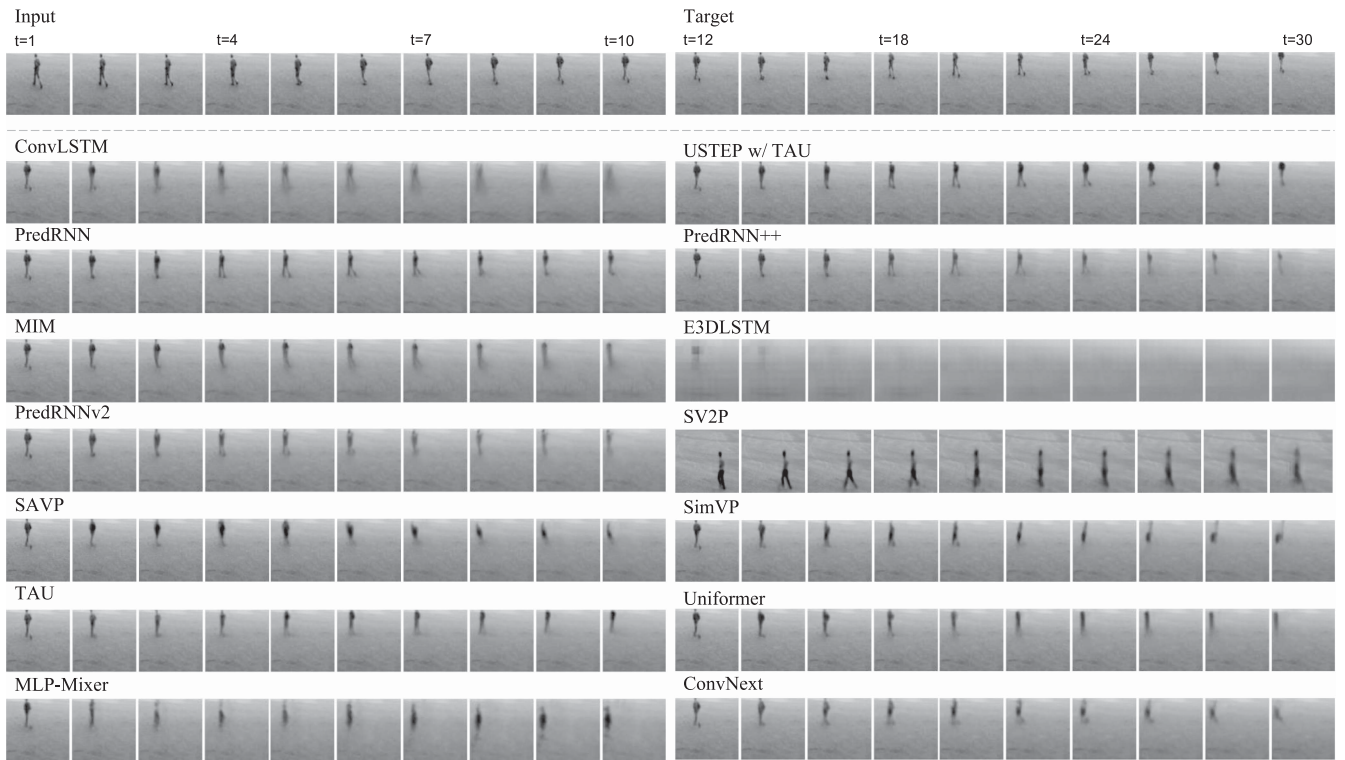
Fig. 6. The qualitative visualization on KTH.

Uniformer, SimVP, and TAU as the basic modules, USTEP consistently improves their baseline models across a range of evaluation metrics, showcasing its robustness and versatility in this setting. For example, USTEP w/TAU achieves an SSIM of 0.9165 and an FVD of 393.13, outperforming both its recurrent-free and recurrent-based counterparts.

To push the boundaries of predictive modeling further, we extended the challenge on the KTH dataset to encompass a longer prediction horizon, specifically from 10 observed frames to forecasting 40 future frames. The results of this ambitious undertaking are meticulously compiled in Table IV. USTEP w/TAU not only maintains but also amplifies its lead over competing models, delivering outstanding performance across a suite of evaluation metrics. This impressive feat underscores USTEP's exceptional capability to generate long sequences of future frames with high fidelity, thereby affirming its robustness and versatility for a broad spectrum of real-world scenarios. In particular, USTEP w/TAU achieves an MSE of 54.68, which is significantly lower than the baseline models, and an SSIM of 0.8832, indicating superior structural similarity in the predicted frames. Furthermore, USTEP w/TAU also demonstrates a notable reduction in FVD scores, highlighting its enhanced ability to produce visually coherent and temporally consistent predictions. These improvements are not limited to one configuration; USTEP variants with Uniformer and SimVP also show considerable improvements, showcasing the framework's flexibility and effectiveness across different underlying architectures.

Such comprehensive evaluations serve not only to benchmark the current state of spatiotemporal predictive modeling but also to illuminate the path forward. While traditional recurrent-based methods like PredRNN struggle to maintain performance over longer prediction horizons, and recurrent-free methods like SimVP tend to overlook short-term dependencies, USTEP manages to harness the strengths of both, setting a new standard in scenarios demanding the prediction of extended future sequences.

C. Reduced Frame Task

Within the scope of the Reduced Frame Task, our objective was to rigorously evaluate the model's adeptness at learning from a limited set of observed frames and its effectiveness in minimizing accumulated prediction errors. This evaluation is critical, as it tests the model's capability to infer future states from sparse temporal data, a common scenario in real-world applications. The Caltech Pedestrian dataset, known for its complexity and real-life relevance, served as the foundation for this assessment. The dataset's diverse and dynamic pedestrian scenes provide a robust testbed for spatio-temporal prediction models. The results from this evaluation are systematically presented in Table V.

Recurrent-based methods such as PredRNN and PredRNN++ exhibit strong overall performance with balanced metrics, indicating their proficiency in handling temporal dependencies. However, these methods tend to have higher computational costs in terms of parameters and FLOPs. Recurrent-free methods like SimVP and TAU demonstrate low computational costs but at the expense of higher prediction errors, suggesting limitations in

TABLE IV
QUANTITATIVE RESULTS ON THE KTH DATASET (10 → 40 FRAMES)

Method	Params	FLOPs	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓
ConvLSTM	14.9	2312.0	63.60	512.2	0.8677	25.42	36.98	1893.45
PredRNN	23.6	4730.0	61.42	499.0	0.8753	25.49	34.26	1711.06
PredRNN++	38.3	7032.0	74.21	554.1	0.8615	24.53	38.44	1482.15
MIM	39.8	1864.0	60.92	520.3	0.8645	25.42	35.14	1578.64
E3DLSTM	53.5	498.0	192.83	1124.5	0.7843	19.87	68.43	3995.82
PredRNNv2	23.6	4757.0	70.27	516.7	0.8554	25.12	37.43	1500.34
SV2P	**8.3**	638.0	85.14	907.1	0.8011	24.08	38.24	1681.09
SAVP	17.6	1084.0	156.43	1198.6	0.8042	24.32	39.90	1691.11
MLP-Mixer	20.3	<u>133.2</u>	81.44	681.0	0.8397	23.36	42.55	2148.75
ConvNeXt	12.5	127.8	58.57	502.1	0.8760	25.59	36.77	1648.97
Uniformer	<u>11.8</u>	156.6	63.04	561.0	0.8601	24.98	39.32	1543.23
SimVP	12.2	**125.6**	68.63	603.5	0.8555	25.10	39.23	1501.04
TAU	15.0	147.6	62.22	511.1	0.8689	25.47	35.77	1500.63
USTEP w/Uniformer	12.0$_{(+1.69\%)}$	202.0$_{(+28.99\%)}$	<u>54.84</u>$_{(-8.20)}$	<u>440.5</u>$_{(-120.5)}$	<u>0.8798</u>$_{(+0.0197)}$	<u>26.10</u>$_{(+1.12)}$	<u>32.45</u>$_{(-6.87)}$	<u>945.24</u>$_{(-597.99)}$
USTEP w/SimVP	16.4$_{(+34.43\%)}$	274.0$_{(+118.15\%)}$	55.23$_{(-13.40)}$	445.6$_{(-157.9)}$	0.8783$_{(+0.0228)}$	26.05$_{(+0.95)}$	32.81$_{(-6.42)}$	1095.72$_{(-405.32)}$
USTEP w/TAU	12.8$_{(-14.67\%)}$	214.0$_{(+44.99\%)}$	**54.68**$_{(-7.54)}$	**433.2**$_{(-77.9)}$	**0.8832**$_{(+0.0143)}$	**26.22**$_{(+0.75)}$	**31.28**$_{(-4.49)}$	**849.16**$_{(-651.47)}$

The results of the USTEP series that underperform the default are in blue while outperforming those are in red.

TABLE V
QUANTITATIVE RESULTS ON THE CALTECH PEDESTRIAN DATASET (10 → 1 FRAME)

Method	Params	FLOPs	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
ConvLSTM	15.0	595.0	139.6	1583.3	0.9345	27.46	8.58
PredRNN	23.7	1216.0	130.4	1525.5	0.9374	27.81	7.40
PredRNN++	38.5	1803.0	125.5	1453.2	0.9433	28.02	13.21
MIM	49.2	1858.0	125.1	1464.0	0.9409	28.10	6.35
E3DLSTM	54.9	1004.0	200.6	1946.2	0.9047	25.45	12.60
PredRNNv2	23.8	1223.0	147.8	1610.5	0.9330	27.12	8.92
MLP-Mixer	22.2	83.5	207.9	1835.9	0.9133	26.29	7.75
ConvNeXt	12.5	80.2	146.8	1630.0	0.9336	27.19	6.99
Uniformer	11.8	104.0	135.9	1534.2	0.9393	27.66	6.87
SimVP	8.6	**60.6**	160.2	1690.8	0.9338	26.81	6.76
TAU	15.0	92.5	131.1	1507.8	0.9456	27.83	<u>5.49</u>
USTEP w/Uniformer	**4.7**$_{(-60.17\%)}$	<u>79.6</u>$_{(-23.46\%)}$	<u>123.8</u>$_{(-12.1)}$	<u>1421.2</u>$_{(-113.0)}$	<u>0.9471</u>$_{(+0.0078)}$	<u>28.32</u>$_{(+0.66)}$	5.50$_{(-1.37)}$
USTEP w/SimVP	6.3$_{(-26.74\%)}$	105.0$_{(+73.27\%)}$	144.6$_{(-15.6)}$	1532.0$_{(-158.8)}$	0.9465$_{(+0.0127)}$	28.09$_{(+1.28)}$	5.65$_{(-1.11)}$
USTEP w/TAU	<u>5.1</u>$_{(-66.00\%)}$	85.3$_{(-7.78\%)}$	**123.6**$_{(-7.5)}$	**1407.9**$_{(-99.9)}$	**0.9477**$_{(+0.0021)}$	**28.37**$_{(+0.54)}$	**4.94**$_{(-0.55)}$

The results of the USTEP series that underperform the default are in blue while outperforming those are in red.

capturing detailed temporal dependencies. In contrast, USTEP w/TAU delivers the best overall performance, achieving the lowest MSE and MAE, highest SSIM and PSNR, and lowest LPIPS, all while significantly reducing computational requirements. This highlights USTEP's ability to balance accuracy and efficiency effectively. Additionally, USTEP variants with Uniformer and SimVP also exhibit notable improvements over their baseline models, demonstrating significant reductions in parameters and FLOPs while maintaining or enhancing performance metrics. These results underscore the efficacy of the USTEP framework in integrating the strengths of both recurrent-based and recurrent-free methods.

The SEVIR dataset is particularly challenging due to the highly dynamic and irregular nature of meteorological events. As shown in Table VI, recurrent-based methods like PredRNN++ achieved strong results, with the highest Critical Success Index (CSI) among non-USTEP models, but at high computational overhead cost. Recurrent-free methods such as SimVP obtained exceptional computational efficiency, with the lowest parameter and FLOP counts, but lagged behind in CSI.

TABLE VI
QUANTITATIVE RESULTS ON THE SEVIR DATASET ($13 \rightarrow 12$ FRAMES) AND UCF SPORTS ($4 \rightarrow 1$ FRAME)

Method	SEVIR				UCF Sports			
	Params	FLOPs	CSI↑	RMSE↓	Params	FLOPs	PSNR↑	LPIPS↓
ConvLSTM	17.4	3097.0	0.4082	13.01	15.0	4017.0	25.90	29.33
PredRNN	27.9	6170.0	0.4288	12.82	23.8	8208.0	27.17	28.15
PredRNN++	42.6	8706.0	0.4312	12.64	38.5	12171.0	27.26	26.80
Uniformer	19.6	408.0	0.3797	13.87	38.7	**223.0**	26.63	28.78
SimVP	12.8	**171.0**	0.3959	12.66	37.9	226.0	27.96	24.99
TAU	20.5	241.0	0.3941	12.73	38.2	236.0	28.10	24.55
USTEP w/Uniformer	**12.2**$_{(-37.75\%)}$	253.0$_{(-37.99\%)}$	0.4091$_{(+0.0294)}$	13.06$_{(-0.81)}$	**5.6**$_{(-85.53\%)}$	235.0$_{(+5.38\%)}$	27.32$_{(+0.69)}$	26.91$_{(-1.87)}$
USTEP w/SimVP	13.1$_{(+2.34\%)}$	267.0$_{(+56.14\%)}$	0.4289$_{(+0.0330)}$	12.62$_{(-0.04)}$	6.1$_{(-83.91\%)}$	251.0$_{(+11.06\%)}$	28.01$_{(+0.05)}$	24.77$_{(-0.22)}$
USTEP w/TAU	13.5$_{(-34.15\%)}$	277.0$_{(+14.93\%)}$	**0.4366**$_{(+0.0425)}$	**12.10**$_{(-0.63)}$	6.3$_{(-83.51\%)}$	262.0$_{(+11.12\%)}$	**28.48**$_{(+0.38)}$	**24.23**$_{(-0.32)}$

The results of the USTEP series that underperform the default are in blue while outperforming those are in red.

TABLE VII
ABLATION STUDY ON THE INFLUENCE OF DIFFERENT DESIGN CHOICES ON THE MOVING MNIST DATASET

Method	Moving MNIST							
	Params	FLOPs	Training time	Inference time	MSE↓	MAE↓	SSIM↑	PSNR↑
USTEP ($\Delta t = 5$)	18.9	17.7	93	58	**21.84**	63.21	**0.9538**	24.06
USTEP ($\Delta t = 1$)	18.8	52.5	600	1052	31.94	71.72	0.9416	23.36
USTEP ($\Delta t = 2$)	18.3	30.8	282	293	24.62	**62.30**	0.9525	**24.46**
USTEP ($\Delta t = 10$)	19.1	13.4	64	35	25.13	74.31	0.9440	23.02
USTEP ($\Delta t = 1, \Delta T = 1$)	18.5	89.8	863	1449	37.21	86.51	0.9240	22.21
w/o cross segment	17.4	13.1	82	52	24.01	67.65	0.9489	23.57
TAU	44.7	16.0	63	39	24.60	71.93	0.9454	23.19
PredRNN++	38.6	171.7	511	373	22.06	69.58	0.9509	23.65

(Δt is default to 10). Training time is measured in seconds per epoch, while inference time is measured in seconds of the test set.

The TAU model provided a balance between accuracy and efficiency, achieving competitive results. The USTEP framework outperformed baselines on the SEVIR dataset. USTEP w/TAU delivered the best performance, achieving the highest CSI and the lowest RMSE, while reducing parameters by 34.15% . The UCF Sports dataset was used to evaluate the model's ability to predict human actions in dynamic scenes. USTEP demonstrated its adaptability and superior performance on this dataset. USTEP w/TAU achieved the best overall results, with the highest PSNR and the lowest LPIPS, while reducing parameters by 83.51% compared to recurrent-based methods.

D. Ablation Study

To unravel the nuances that underpin the performance of USTEP, an in-depth ablation study was undertaken. This meticulous examination was directed towards understanding the repercussions of various design decisions, with a particular focus on the temporal stride (Δt) and the integration of a cross-segment mechanism within the model's architecture. For convenience, we refer to USTEP w/TAU as simply USTEP. The findings from this study succinctly encapsulated in Table VII, shed light on the intricate balance between computational efficiency and predictive accuracy.

The optimal performance is achieved when Δt is set to 5, where USTEP records an MSE of 21.84 and an SSIM of 0.9538, marking the zenith of its predictive accuracy. This particular configuration of Δt underscores a harmonious balance between capturing essential temporal dynamics and maintaining computational pragmatism. With considerable performance improvements over TAU, USTEP only marginally sacrifices training and inference time while still remaining significantly faster than its recurrent-based counterparts. However, diminishing the stride (Δt) heightens the model's computational demand, evidenced by an uptick in FLOPs, attributed to the increment in computational occurrences necessitated by shorter temporal intervals.

The role of the cross-segment mechanism emerges as pivotal within this contextual framework. Its presence is instrumental, acting as a crucial part of USTEP's architecture that significantly enhances spatiotemporal predictive performance. Disabling this cross-segment mechanism results in noticeable performance degradation, emphasizing its criticality in facilitating a nuanced interplay between different temporal segments, thereby enriching the model's predictive depth and accuracy.

We compare the convergence curves between USTEP and representative recurrent-based method PredRNN++ and recurrent-free method TAU in Fig. 7. While TAU benefits from fast convergence in the earlier epochs, USTEP gradually catches
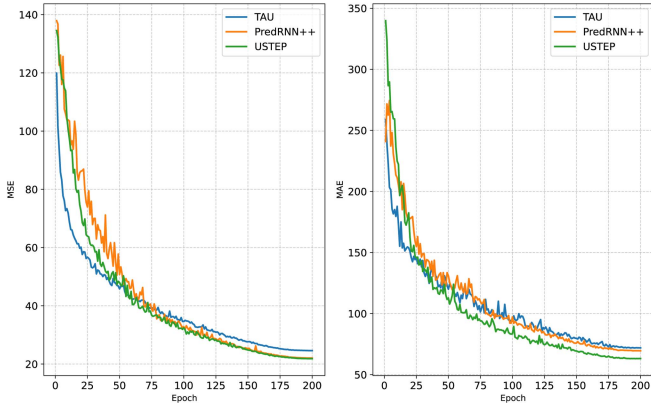
Fig. 7.   Convergence comparison between USTEP and strong baselines, PredRNN++ (recurrent-based) and TAU (recurrent-free).

Fig. 9.   Representative failure cases of USTEP on complex variants of Moving MNIST.
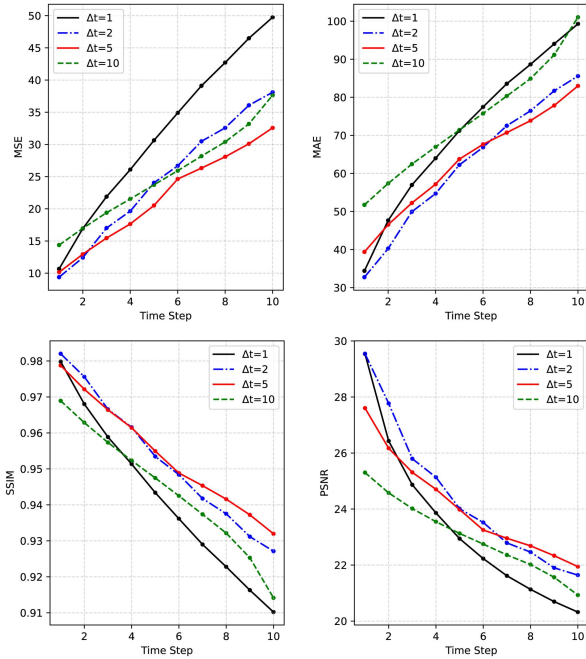

Fig. 8.   Frame-wise comparison in MSE, MAE, SSIM and PSNR metrics. For MSE and MAE metrics, lower values are preferable. For SSIM and PSNR metrics, higher values are more desirable.

up and eventually surpasses TAU. PredRNN++ exhibits the slowest convergence compared to USTEP. Moreover, while USTEP achieves a similar MSE to PredRNN++, it significantly outperforms PredRNN++ in terms of MAE. This suggests that USTEP produces more accurate and visually coherent predicted frames, effectively reducing larger prediction errors. USTEP not only offers a compatible solution that bridges the gap between recurrent-based and recurrent-free models but also provides strong overall performance, making it a robust choice for spatio-temporal predictive learning tasks.

Fig. 8 delineates that $\Delta t = 1$ tends to overemphasize local information, potentially leading to a lack of holistic understanding. In contrast, $\Delta t = 10$ appears to overly prioritize
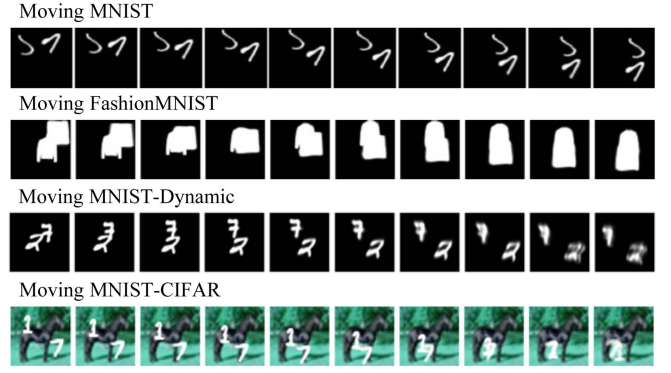
global information, possibly at the expense of missing finer, localized details. These insights underline the importance of choosing an appropriate Δt in USTEP to balance local and global temporal considerations, ensuring the holistic integrity of the learned features and predictions.

E. Empirical Analysis and Practical Guideline

Our objective is to integrate the strengths of both recurrent-based and -free approaches for spatiotemporal predictive learning. Using a fixed time step Δt is a simple yet effective way to achieve this. As we have set ΔT as the length of the input sequence in Definition 3.3, the only hyperparameter left to be tuned is Δt. Empirically, we recommend using $\Delta t = 5$ for long videos $(T + T' > 10)$ and $\Delta t = 2$ for short videos $(T + T' \leq 10)$. As indicated in Table VIII, minor changes were observed on KTH and WeatherBench. The other datasets remained unchanged in terms of metrics since their optimal Δt values were already aligned with our principle. In essence, our empirical principle serves as a foundational guideline for selecting the temporal stride Δt when processing video data of varying lengths, guarantees good performance across a wide range of applications.

In practice, Δt can also be adjusted based on the frame rate of the video data. For low frame rates (e.g., 1-5 FPS), a smaller Δt is preferable to ensure no critical temporal information is skipped. For higher frame rates (e.g., 30 FPS or more), a larger Δt can help improve computational efficiency while maintaining performance. Our empirical findings suggest that setting $\Delta t = 5$ for long videos and $\Delta t = 2$ for short videos provides a default configuration.

F. Analysis Between USTEP and Representative Baselines

Recurrent-based models like PredRNN, PredRNN++, and ConvLSTM excel at modeling temporal dependencies by processing frames sequentially, but they suffer from the following limitations:

- *Redundant Frame-by-Frame Processing:* Frame-by-frame processing leads to significant computational overhead,

TABLE VIII
EMPIRICAL Δt EFFECTS

Dataset	Δt	Length	MSE↓	MAE↓	SSIM$_{(\times 10^{-2})}$ ↑	PSNR↑
MMNIST	5	20	21.84$_{(+0.00)}$	63.21$_{(+0.00)}$	95.38$_{(+0.00)}$	24.06$_{(+0.00)}$
KTH	5	30	39.91$_{(+0.36)}$	355.3$_{(-9.6)}$	91.55$_{(-0.10)}$	28.99$_{(+0.01)}$
Caltech	5	11	123.6$_{(+0.00)}$	1407.9$_{(+0.00)}$	94.77$_{(+0.00)}$	28.37$_{(+0.00)}$
WeatherBench	5	24	1.14$_{(-0.01)}$	0.69$_{(+0.04)}$	-	-
Human	2	8	109.5$_{(+0.00)}$	1380.5$_{(+0.00)}$	98.45$_{(+0.00)}$	34.35$_{(+0.00)}$

The results that underperform the default are in blue while outperforming those are in red.

TABLE IX
RUNTIME COMPARISON ON THE KTH (10 → 40 FRAMES) DATASET

Device	Model	Training Time	Inference Time
V100	TAU	188	46
	PredRNN++	6360	1551
	USTEP	221	54
A100	TAU	101	25
	PredRNN++	3405	782
	USTEP	121	31
CPU	TAU	10528	2510
	PredRNN++	362520	88422
	USTEP	12155	2968

Training time is measured in seconds per epoch, while inference time is measured in seconds of the test set.

as short-term spatio-temporal redundancies are repeatedly processed. This inefficiency becomes even more pronounced with longer input sequences. For example, in Table II (Moving MNIST), the FLOPs of ConvLSTM, PredRNN, and PredRNN++ are 56.8, 116.0, and 171.7 GFLOPs, respectively. This shows how the increased complexity of recurrent-based models results in higher computational costs, despite moderate improvements in performance. PredRNN++ achieves better MSE (22.06) compared to PredRNN (23.97) and ConvLSTM (29.80), but the improvement comes at a significant cost in terms of computational efficiency.

- *Complex Module Design:* PredRNN introduces spatio-temporal memory flow, enhancing its ability to capture transitions between frames. PredRNN++ further builds on this by designing more complicated recurrent modules, but the added complexity doesn't always translate to better performance. For instance, in Table IV, PredRNN++'s FLOPs (7032 GFLOPs) are much higher than ConvLSTM (2312 GFLOPs) and PredRNN (4730 GFLOPs). However, its MSE (74.21) is worse than PredRNN (61.42) and even ConvLSTM in some cases. This suggests that the additional complexity in PredRNN++ does not always yield better results, particularly for long-term predictions.

Recurrent-free models such as SimVP and TAU offer an efficient alternative by processing sequences in parallel. However, their limitations include:

- *Loss of Fine-Grained Temporal Interactions:* SimVP processes sequences holistically, focusing on global patterns but neglecting short-term temporal interactions between consecutive frames. This makes it less effective for tasks requiring precise frame-by-frame modeling. For instance, on Moving MNIST (Table II), SimVP achieves an MSE of 32.15, which is significantly worse than PredRNN (23.97) and PredRNN++ (22.06). Its performance gap demonstrates its inability to capture fine-grained dependencies.
- *Efficiency versus Performance Trade-off:* TAU simplifies SimVP by replacing the middle U-Net with plain temporal attention units. TAU demonstrates exceptional efficiency compared to recurrent-based models. For example, on Table IV, TAU achieves a comparable MSE (62.22) with only 147 GFLOPs, whereas PredRNN requires 4730 GFLOPs, and PredRNN++ requires 7032 GFLOPs. However, TAU's focus on global context still limits its ability to model short-term dynamics as effectively as recurrent-based models. While TAU strikes a better balance, it still prioritizes long-term global modeling at the expense of short-term interactions.

USTEP combines the strengths of both recurrent-based and recurrent-free models while addressing their individual limitations. It achieves this through its hierarchical dual-scale framework, which integrates micro-temporal (short-term) and macro-temporal (long-term) modeling:

- *Avoids Redundant Frame-by-Frame Processing and Captures Fine-Grained Temporal Interactions:* Unlike recurrent-based methods, USTEP processes long-term dependencies using macro-temporal segments in a recurrent-free manner, reducing computational overhead while maintaining global context. By incorporating micro-temporal segments, USTEP explicitly models short-term dependencies that recurrent-free models like SimVP and TAU neglect.
- *Dynamic Integration of Temporal Scales:* USTEP employs a dual-gate mechanism that dynamically integrates information from both micro- and macro-temporal scales. USTEP achieves superior performance without incurring the high computational costs of recurrent-based models. For example: On Moving MNIST (Table II), USTEP w/TAU achieves an MSE of 21.84 with 17.7 GFLOPs, outperforming PredRNN++ (22.06 MSE with 171.7 GFLOPs) at only about 10% additional computational cost compared to TAU. On KTH (Table IV), USTEP w/TAU achieves the best performance with an MSE of 54.68 at 214 GFLOPs, significantly outperforming PredRNN (MSE 61.42, 4730 GFLOPs) and PredRNN++ (MSE 74.21, 7032 GFLOPs). USTEP delivers this superior performance while using over 20× fewer FLOPs than recurrent-based models.

USTEP bridges the gap between recurrent-based and recurrent-free methods by combining their strengths and addressing their limitations. It avoids the inefficiencies of

frame-by-frame processing in recurrent-based models and overcomes the lack of fine-grained temporal modeling in recurrent-free models. This unified approach ensures that USTEP can handle both short-term and long-term dependencies dynamically and efficiently. As demonstrated in Section V, USTEP consistently achieves superior performance across benchmarks while maintaining a balanced computational cost, making it an effective and scalable solution for spatio-temporal predictive learning.

G. Runtime Comparison on Different Devices

To further emphasize the computational efficiency of our proposed framework, we conduct a detailed runtime comparison across different hardware platforms, using the KTH dataset with the extended frame prediction task ($10 \rightarrow 40$ frames) as a benchmark. This setup not only provides a challenging long-horizon forecasting scenario but also reveals the practical deployment costs across a range of devices, including high-performance GPUs (NVIDIA V100 32GB, NVIDIA A100 80GB) and a representative CPU platform (Intel(R) Xeon(R) Platinum 8358 @ 2.60GHz). Table IX presents the training and inference times of USTEP, PredRNN++, and TAU on different platforms. On both GPU platforms, USTEP achieves nearly the same efficiency as TAU. Moreover, USTEP demonstrates significantly faster performance compared to PredRNN++.

VI. CONCLUSION AND LIMITATION

This paper introduced USTEP, a novel paradigm for spatiotemporal prediction tasks, thoughtfully architected to unify the strengths of both recurrent-based and recurrent-free models. USTEP operates under a novel paradigm that offers a comprehensive view of spatiotemporal dynamics, facilitating a nuanced understanding and representation of intricate temporal patterns and dependencies. USTEP has proven its mettle across a variety of spatiotemporal tasks, demonstrating exceptional adaptability and superior performance in diverse contexts. It meticulously integrates local and global spatiotemporal information, providing a unified perspective that enhances its performance.

To better understand USTEP's limitations, we conducted an in-depth failure analysis by designing three extended variants of the Moving MNIST that introduce controlled complexities: (i) Moving FashionMNIST: we replace digit characters with fashion items, increasing object variability and deformability; (ii) Moving MNIST-Dynamic: we introduce random Gaussian perturbations to the speed of each object, simulating non-uniform motion patterns; (iii) Moving MNIST-CIFAR: we use natural image backgrounds sampled from the CIFAR-10, introducing substantial background clutter and color variation. These variants are specifically constructed to test USTEP under conditions of occlusion, fast or irregular motion, and complex visual scenes. As illustrated in Fig. 9, USTEP exhibits decreased predictions in these cases, particularly when object motion is highly erratic or when the foreground and background share similar textures or color distributions. In such settings, the model may produce blurred or temporally inconsistent predictions, revealing a sensitivity to dynamic and context-heavy environments. We believe these findings provide valuable insight into USTEP's failure

modes. In future work, we aim to address these challenges by incorporating adaptive motion modeling and stronger attention mechanisms for background-foreground separation.

REFERENCES

[1] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 64–72.

[2] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 515–524.

[3] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4114–4124.

[4] K. Greff et al., "Multi-object representation learning with iterative variational inference," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2424–2433.

[5] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4402–4412.

[6] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2207–2217.

[7] L. Castrejon, N. Ballas, and A. Courville, "Improved conditional VRNNs for video prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7608–7617.

[8] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[9] L. Shen, Z. Li, and J. Kwok, "Timeseries anomaly detection using temporal hierarchical one-class network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 13016–13026.

[10] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[11] Z. Gao et al., "Earthformer: Exploring space-time transformers for earth system forecasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 25390–25403.

[12] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "WeatherBench: A benchmark data set for data-driven weather forecasting," *J. Adv. Model. Earth Syst.*, vol. 12, no. 11, 2020, Art. no. e2020MS002203.

[13] S. Ning et al., "MIMO is all you need: A strong multi-in-multi-out baseline for video prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1975–1983.

[14] M. Seo, H. Lee, D. Kim, and J. Seo, "Implicit stacked autoregressive model for video prediction," 2023, *arXiv:2303.07849*.

[15] S. Fang, Q. Zhang, G. Meng, S. Xiang, and C. Pan, "GSTNet: Global spatial-temporal network for traffic flow prediction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 2286–2293.

[16] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order nonstationarity from spatiotemporal dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9154–9162.

[17] A. Lerer, S. Gross, and R. Fergus, "Learning physical intuition of block towers by example," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 430–438.

[18] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 879–888.

[19] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[20] Y. Wang, Z. Gao, M. Long, J. Wang, and S. Y. Philip, "PredRNN++: Towards a resolution of the deep-in-time Dilemma in spatiotemporal predictive learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5123–5132.

[21] Y. Wang et al., "PredRNN: A recurrent neural network for spatiotemporal predictive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2208–2225, Feb. 2023.

[22] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3D LSTM: A model for video prediction and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[23] B. Jin et al., "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4554–4563.

[24] M. Babaeizadeh, M. T. Saffar, S. Nair, S. Levine, C. Finn, and D. Erhan, "FitVid: Overfitting in pixel-level video prediction," 2021, *arXiv:2106.13195*.

[25] C. Tan et al., "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18770–18782.

[26] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "SimVP: Simpler yet better video prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3170–3180.

[27] C. Tan, Z. Gao, S. Li, and S. Z. Li, "SimVP: Towards simple yet powerful spatiotemporal predictive learning," 2022, *arXiv:2211.12509*.

[28] C. Tan et al., "OpenSTL: A comprehensive benchmark of spatio-temporal predictive learning," 2023, *arXiv:2306.11249*.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] A. S. Marc'Aurelio Ranzato, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: A baseline for generative models of natural videos," 2014, *arXiv:1412.6604*.

[31] Z. Xu, Y. Wang, M. Long, J. Wang, and M. KLiss, "PredCNN: Predictive learning with cascade convolutions," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2940–2947.

[32] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "TrajectoryCNN: A new spatio-temporal feature learning network for human motion prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2133–2146, Jun. 2021.

[33] B. Peng et al., "RWKV: Reinventing RNNs for the transformer era," 2023, *arXiv:2305.13048*.

[34] Y. Sun et al., "Retentive network: A successor to transformer for large language models," 2023, *arXiv:2307.08621*.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[36] X. Ma et al., "Mega: Moving average equipped gated attention," 2022, *arXiv:2209.10655*.

[37] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11474–11484.

[38] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

[39] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[40] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[41] W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, "Efficient and information-preserving future frame prediction and beyond," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–12.

[42] M. Veillette, S. Samsi, and C. Mattioli, "SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 22009–22019.

[43] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[44] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, pp. 32–36.

[45] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–11.

[46] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–14.

[47] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," 2018, *arXiv: 1804.01523*.

[48] K. Li et al., "Uniformer: Unifying convolution and self-attention for visual recognition," 2022, *arXiv:2201.09450*.

[49] I. O. Tolstikhin et al., "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.

[50] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[52] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "FVD: A new metric for video generation," 2019. [Online]. Available: https://openreview.net/forum?id=rylgEULtdN

Cheng Tan (Graduate Student Member, IEEE) is currently working toward the PhD degree with Zhejiang University (ZJU) and Westlake University, Hangzhou, China, supervised by Chair Professor Stan Z. Li. His main research interests include self-supervised learning and spatiotemporal learning. He has published several papers in top conferences and journals such as NeurIPS, ICML, ICLR, CVPR, ECCV, AAAI, ACM MM, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, etc.

Jue Wang is currently working as research assistant with the CAIRI Lab, Westlake University, supervised by Chair Professor Stan Z. Li. His main research interests include self-supervised learning.

Zhangyang Gao received the BS degree from Nanjing University, Nanjing, China, in 2021. He is currently working toward the PhD degree with Zhejiang University (ZJU) and Westlake University, Hangzhou, China. His main research interests include clustering, unsupervised video tasks, and deep learning. He has published several papers in top conferences and journals such as CVPR, ICML, ICLR, NeurIPS, AAAI, *IEEE Transactions on Knowledge and Data Engineering*, etc.

Siyuan Li (Graduate Student Member, IEEE) received the BS degree from the School of Automation, Central South University, Changsha, China, in 2020. He is currently working toward the PhD degree with Zhejiang University (ZJU) and Westlake University, Hangzhou, China. His main research interests include clustering, unsupervised video tasks, and deep learning. He has published several papers in top conferences and journals, such as ICML, NeurIPS, ICLR, CVPR, ECCV, IJCAI, BMVC, etc.

Stan Z. Li (Fellow, IEEE) received the BEng degree from Hunan University, China, the MEng degree from the National University of Defense Technology, China, and the PhD degree from the University of Surrey, U.K. He is currently a professor and the director with the Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). From 2000 to 2004, he worked with Microsoft Research Asia, as a researcher. Prior to that, he was an associate professor with Nanyang Technological University, Singapore. He has published more than 200 papers in international journals and conferences and authored and edited eight books. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He was elevated to a fellow of the IEEE, for his contributions to the fields of face recognition, pattern recognition, and computer vision. He served as the program co-chair for the International Conference on Biometrics in 2007 and 2009, and has been involved in organizing other international conferences and workshops in the fields of his research interest.