UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding

Kunchang Li\textsuperscript{1,2,3*} Yali Wang\textsuperscript{1,3†} Yinan He\textsuperscript{3} Yizhuo Li\textsuperscript{4,3*} Yi Wang\textsuperscript{3} Limin Wang\textsuperscript{5,3} Yu Qiao\textsuperscript{3,1†}

\textsuperscript{1}Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
\textsuperscript{2}University of Chinese Academy of Sciences \textsuperscript{3}Shanghai AI Laboratory \textsuperscript{4}The University of Hong Kong
\textsuperscript{5}State Key Laboratory for Novel Software Technology, Nanjing University

Code & Models: https://github.com/OpenGVLab/UniFormerV2

Abstract

The prolific performances of Vision Transformers (ViTs) in image tasks have prompted research into adapting the image ViTs for video tasks. However, the substantial gap between image and video impedes the spatiotemporal learning of these image-pretrained models. Though video-specialized models like UniFormer can transfer to the video domain more seamlessly, their unique architectures require prolonged image pretraining, limiting the scalability. Given the emergence of powerful open-source image ViTs, we propose unlocking their potential for video understanding with efficient UniFormer designs. We call the resulting model UniFormerV2, since it inherits the concise style of the UniFormer block, while redesigning local and global relation aggregators that seamlessly integrate advantages from both ViTs and UniFormer. Our UniFormerV2 achieves state-of-the-art performances on 8 popular video benchmarks, including scene-related Kinetics-400/600/700, heterogeneous Moments in Time, temporal-related Something-Something V1/V2, and untrimmed ActivityNet and HACS. It is noteworthy that to the best of our knowledge, UniFormerV2 is the first to elicit 90% top-1 accuracy on Kinetics-400.

1. Introduction

The triumph of transformer-based language foundation models [16, 51, 5] has resulted in the swift growth of image foundation models [18, 24, 50, 73], which have been meticulously trained on massive web datasets with rich supervision, such as image-text contrastive learning [50, 30] and mask image modeling [24, 3]. The resulting Vision Transformers (ViTs) exhibit exceptional generalization capacity for a range of image tasks [43, 12, 53], motivating researchers to explore their applications for video tasks.

*Interns at Shanghai AI Laboratory. †Corresponding authors.
ViT-based video learners surpass CNNs by a considerable margin on traditional scene-related benchmarks [32, 9, 10], which can be recognized easily by a single frame. However, when faced with complex temporal-related tasks [22], they perform much worse than CNN-based ones [34, 62]. The substantial domain gap between image and video presents a challenge to adapt image ViTs for video understanding.

Another prevalent paradigm is to design specialized ViTs [42, 37, 35], which can be effortlessly transferred to the video domain via simple technique, i.e., inverting spatial convolution or attention to spatiotemporal ones. In the advanced UniFormer [35], the authors unify convolution and self-attention as Multi-Head Relation Aggregator (MHRA) in a transformer format. By modeling local and global relations respectively in shallow and deep layers, it can not only handle both scene-related and temporal-related tasks effectively, but also significantly reduce the computation burden. However, as a unique architecture, UniFormer lacks image pretraining as a starting point. To obtain a robust visual representation, it has to go through prolonged pretraining on images before finetuning on videos, which makes it difficult to scale up. Considering the emergence of powerful open-source image ViTs [66, 3, 50], a natural question arises: Can we unlock the potential of image ViTs for video understanding with an efficient UniFormer design?

In this paper, we propose a simple yet effective paradigm for constructing powerful video networks, by arming the image-pretrained ViTs with efficient UniFormer designs (see Figure 1). We call the resulting model UniFormerV2, since it inherits the concise style of UniFormer but equips local and global UniBlocks with new MHRA. In the local UniBlock, we incorporate a local temporal MHRA before the spatial ViT block. Thus we can largely reduce temporal redundancy and leverage the well-pretrained ViT block, for learning local spatiotemporal representation effectively. As for the global UniBlock, we introduce a query-based cross MHRA. Unlike the costly global MHRA in the original UniFormer, our cross MHRA can summarize all the spatiotemporal tokens into a video token, for learning global spatiotemporal representation efficiently. Finally, we reorganize local and global UniBlocks as a multi-stage fusion architecture, which can adaptively integrate multi-scale spatiotemporal representation to capture complex dynamics.

We apply our paradigm on ViTs that are pretrained on three popular supervision, including supervised learning [55, 56], contrastive learning [50], and mask image modeling [24, 3]. Our results reveal that all enhanced models exhibit superior performance compared to previous ViT-based approaches, showcasing the generic nature of our UniFormerV2. In addition, we have constructed a compact Kinetics-710 benchmark, combining the action classes of Kinetics-400/600/700, and have removed repeated and leaked videos in the training sets of these benchmarks for enhanced fairness. As a result, the number of training videos has been reduced from 1.14M to 0.66M. After training on K710, our model can simply achieve higher accuracy on K400/600/700 via only 5-epoch finetuning.

To verify the robustness of our approach, we conduct experiments on 8 large-scale video benchmarks as shown in Figure 1, including scene-related datasets (i.e., Kinetics-400/600/700 [32, 9, 10], a heterogeneous dataset that contains complex inter-class and inter-class variation (i.e., Moments in Time [44]), temporal-related datasets (i.e., Something-Something V1/V2 [22]), and untrimmed datasets (i.e., ActivityNet [25] and HACS [78]). Our UniFormerV2 based on CLIP-ViT [50] achieves state-of-the-art results on all the benchmarks. It is worth mentioning that our model is the first to elicit a top-1 accuracy of 90.0% on Kinetics-400, to the best of our knowledge.

2. Related Works

**Vision Transformer.** Following the groundbreaking success of Transformer in NLP [60, 16], Vision Transformer (ViT) [18] has shown great promise in a variety of visual tasks, including object detection [7, 81], semantic segmentation [70, 13], low-level image processing [38, 15], action recognition [4, 1, 72], temporal localization [76, 61] and multi-modality learning [50, 64]. To further enhance the efficiency and effectiveness of ViT, researchers have explored various methods for modeling locality, including multi-scale architectures [65, 19], local window [41], early convolution embedding [69, 74] and convolutional position encoding [14, 17]. Alternatively, UniFormer [35] unifies convolution and self-attention as relation aggregator in a transformer manner, thus reducing large local redundancy.

**Video Learning.** 3D Convolutional Neural Networks (CNNs) once played a dominant role in video understanding [57, 11]. However, the optimization of 3D CNNs can be problematic, hence great efforts have been made to factorize 3D convolution in the spatiotemporal dimension [59, 49, 21] or channel dimension [58, 20, 33]. Other advanced methods propose plug-and-play modules to enhance the temporal modeling capability of 2D CNNs [39, 31, 36, 34, 62]. However, due to the restricted local receptive field, CNNs are apt to miss long-range dependencies. The success of global attention [18] motivates researchers to adapt image ViTs for video tasks [4, 45, 77, 1, 6, 48]. To make the video transformer more efficient, prior works introduce hierarchical structure with pooling self-attention [19], local self-attention [42] or unified attention [35]. Though these novel models are adept at temporal modeling, they rely on tiresome image pretraining. In contrast, various well-pretrained ViTs with rich supervision are open-sourced [66, 3, 50]. In this paper, we aim to extend efficient UniFormer designs to ViT, arming it as a strong video learner.
3. Method

3.1. Revisit UniFormer

UniFormer [35] is originally proposed for efficient video understanding. It unifies convolution and self-attention as Multi-Head Relation Aggregator (MHRA) in a transformer format as shown in the bottom-left in Figure 2, along with Dynamic Position Embedding (DPE) and Feed-Forward Network (FFN). Specifically, the DPE is instantiated as 3×3×3 depth-wise spatiotemporal convolution to integrate 3D position information. And the FFN includes two linear layers for pointwise enhancement. Similar with Multi-Head Self-Attention (MHSA) [50], the MHRA learns token relation via multi-head fusion:

\[
R_n(X) = A_n V_n(X), \quad (1)
\]
\[
\text{MHRA}(X) = \text{Concat}(R_1(X); \ldots; R_N(X))U, \quad (2)
\]

where \(R_n(\cdot)\) refers to the relation aggregator in the \(n\)-th head. \(A_n\) is an affinity matrix that describes token relation and \(V_n(\cdot)\) is a linear projection, while \(U \in \mathbb{R}^{C \times C}\) is a learnable fusion matrix. The crucial MHRA flexibly applies local and global spatiotemporal token affinity in the shallow and deep layers, respectively, tackling both video local redundancy and global dependency.

However, like other specialized video backbones [19, 37, 42], UniFormer is difficult to scale up due to the necessity of costly image pretraining. Considering the emergence of powerful image ViTs [66, 3, 50], it is preferable to arm those of costly image pretraining. Considering the specialization of UniFormer [35], we adopt Batch Normalization (BN) [28] before local MHRA, and Layer Normalization (LN) [2] before global MHRA and FFN. Note that GS_MHRA and FFN come from the image-pretrained ViT block. Driven by the architectural insight of UniFormer, we incorporate LT_MHRA to mitigate local temporal redundancy effectively. Hence, the affinity in LT_MHRA is local with a learnable parameter matrix \(a_n \in \mathbb{R}^{t \times 1 \times 1}\) in the temporal tube \(t \times 1 \times 1\),

\[
A_n^{LT}(X_i, X_j) = a_n^{i-j}, \quad \text{where} \quad j \in \Omega_i^{t \times 1 \times 1}. \quad (6)
\]

This allows to efficiently learn the local temporal relation between one token \(X_i\) and other tokens \(X_j\) in the tube. Alternatively, GS_MHRA belongs to the original ViT block. Therefore, the affinity in GS_MHRA refers to a global spatial self-attention in the single frame \(1 \times H \times W\),

\[
A_n^{GS}(X_i, X_j) = \frac{\exp\{Q_n(X_i)^T K_n(X_j)\}}{\sum_{j'} \in \Omega_i^{1 \times H \times W} \exp\{Q_n(X_i)^T K_n(X_{j'})\}}, \quad (7)
\]

where \(Q_n(\cdot)\) and \(K_n(\cdot) \in \mathbb{R}^{L \times C}\) are different linear projections in the \(n\)-th head.

Comparison to UniFormer: In the UniFormer [35], the local token affinity is jointly spatiotemporal, i.e., \(A_n^{local}(X_i, X_j) = a_n^{i-j}\), where \(j\) belongs to a 3D tube \(\Omega_i^{t \times h \times w}\). And the parameter matrix has to learn from scratch, which inevitably increases the training cost. In contrast, the spatiotemporal affinity in our local UniBlock is decomposed as local temporal one \(A_n^{LT}\) in Eq. (6), and global spatial one \(A_n^{GS}\) in Eq. (7). In this case, we can not only leverage the efficient video processing design of UniFormer but also inherit the effective image pretraining of ViT.

Comparison to ST-Adapter: ST-Adapter [47] is motivated by Adapter [27], thus it simply treats temporal depthwise convolution as adaptation and introduces an extra activation function. In contrast, inspired by UniFormer [35], all the spatiotemporal tokens as a global video token. Finally, all tokens with different-level global semantics from multiple stages are fused together to form a discriminative video representation.

3.3. Local UniBlock

To efficiently model temporal dependency upon the well-learned spatial representation, we insert the novel local temporal MHRA before the standard ViT block,

\[
X_T = \text{LT}_\text{MHRA} \left( \text{Norm}(X^\text{in}) \right) + X^\text{in}, \quad (3)
\]
\[
X_S = \text{GS}_\text{MHRA} \left( \text{Norm}(X^T) \right) + X^T, \quad (4)
\]
\[
X_L = \text{FFN} \left( \text{Norm}(X^S) \right) + X^S. \quad (5)
\]

LT_MHRA and GS_MHRA refer to MHRA with local temporal affinity and global spatial affinity respectively. FFN consists of two linear projections separated by GeLU [26]. Additionally, following the normalization in UniFormer [35], we adopt Batch Norm (BN) [28] before local MHRA, and Layer Norm (LN) [2] before global MHRA and FFN. Note that GS_MHRA and FFN come from the image-pretrained ViT block. Driven by the architectural insight of UniFormer, we incorporate LT_MHRA to mitigate local temporal redundancy effectively. Hence, the affinity in LT_MHRA is local with a learnable parameter matrix \(a_n \in \mathbb{R}^{t \times 1 \times 1}\) in the temporal tube \(t \times 1 \times 1\),

\[
A_n^{LT}(X_i, X_j) = a_n^{i-j}, \quad \text{where} \quad j \in \Omega_i^{t \times 1 \times 1}. \quad (6)
\]

This allows to efficiently learn the local temporal relation between one token \(X_i\) and other tokens \(X_j\) in the tube. Alternatively, GS_MHRA belongs to the original ViT block. Therefore, the affinity in GS_MHRA refers to a global spatial self-attention in the single frame \(1 \times H \times W\),

\[
A_n^{GS}(X_i, X_j) = \frac{\exp\{Q_n(X_i)^T K_n(X_j)\}}{\sum_{j'} \in \Omega_i^{1 \times H \times W} \exp\{Q_n(X_i)^T K_n(X_{j'})\}}, \quad (7)
\]

where \(Q_n(\cdot)\) and \(K_n(\cdot) \in \mathbb{R}^{L \times C}\) are different linear projections in the \(n\)-th head.

Comparison to UniFormer: In the UniFormer [35], the local token affinity is jointly spatiotemporal, i.e., \(A_n^{local}(X_i, X_j) = a_n^{i-j}\), where \(j\) belongs to a 3D tube \(\Omega_i^{t \times h \times w}\). And the parameter matrix has to learn from scratch, which inevitably increases the training cost. In contrast, the spatiotemporal affinity in our local UniBlock is decomposed as local temporal one \(A_n^{LT}\) in Eq. (6), and global spatial one \(A_n^{GS}\) in Eq. (7). In this case, we can not only leverage the efficient video processing design of UniFormer but also inherit the effective image pretraining of ViT.

Comparison to ST-Adapter: ST-Adapter [47] is motivated by Adapter [27], thus it simply treats temporal depthwise convolution as adaptation and introduces an extra activation function. In contrast, inspired by UniFormer [35],
we treat temporal depth-wise convolution as a local temporal relation aggregator, thus introducing extra BatchNorm [28] before the first linear projection $V(\cdot)$ without any activation function. As evidenced by Table 2, our local MHRA outperforms ST-Adapter (69.1% vs. 68.0%).

3.4. Global UniBlock

To explicitly conduct long-range dependency modeling on the spatiotemporal scale, we present the global UniBlock as follows,

$$X^C = \text{DPE}(X^L) + X^L, \quad (8)$$

$$X^{ST} = \text{C-MHRA}(\text{Norm}(q), \text{Norm}(X^C)), \quad (9)$$

$$X^G = \text{FFN}(\text{Norm}(X^{ST})) + X^{ST}. \quad (10)$$

Following UniFormer [35], we apply DPE to dynamically integrate 3D position information. Moreover, we redesign the global C_MHRA in cross-attention style to efficiently construct a video representation,

$$R_n^C(q, X) = A_n^C(q, X)V_n(X), \quad (11)$$

$$\text{C-MHRA}(q, X) = \text{Concat}(R_1^C(q, X), \cdots ; R_n^C(q, X))U. \quad (12)$$

$R_n^C(q, \cdot)$ is the cross relation aggregator, which can convert a learnable query $q \in \mathbb{R}^{1 \times C}$ into a video representation, via modeling dependency between $q$ and all the spatiotemporal tokens $X$. First, it computes the cross affinity matrix $A_n^C(q, X)$ to learn relation between $q$ and $X$,

$$A_n^C(q, X^j) = \frac{\exp\{Q_n(q)^T K_n(X^j)\}}{\sum_{j' \in \Omega_{T \times H \times W}} \exp\{Q_n(q)^T K_n(X^j')\}}. \quad (13)$$

Then, it uses the linear projection to transform $X$ as spatiotemporal context $V_n(X)$. Subsequently, it aggregates such context $V_n(X)$ into the learnable query, with guidance of their affinity $A_n^C(q, X)$. Finally, the enhanced query tokens from all the heads are further fused as a final video representation, by linear projection $U \in \mathbb{R}^{C \times C}$. Note the query token is zero-initialized for stable training.

**Comparison to UniFormer**: The global spatiotemporal MHRA present in UniFormer [35] is computationally heavy due to the quadratic complexity it entails. In contrast, our global MHRA in cross-attention style significantly reducing the computation complexity from $O(L^2)$ to $O(L)$, where $L$ is the number of tokens. More importantly, through the learnable query $q$, our global MHRA can adaptively incorporate spatiotemporal context from all $L$ tokens to enhance video recognition. Furthermore, we add the global UniBlock on top of the local UniBlock, extracting multiscale spatiotemporal representations in token form. This design helps strengthen the discriminative video representation without compromising the pretrained architecture.

**Comparison to DETR style**: The methods inspired by DETR [7, 29] incorporate self-attention, cross-attention, and FFN. And they employ multiple queries with identical keys and values in cross-attention. On the other hand, our global block introduces DPE without self-attention. Meanwhile, only one query interacts with keys and values from distinct layers in our cross-attention.

3.5. Multi-Stage Fusion Block

We propose a multi-stage fusion block to integrate all video tokens from each global block as in Figure 3. For simplicity, we denote the $i$-th global block as $X^G_i = G_i(q, X^L_i)$. Given the tokens $X^L_i$ from the local UniBlock, the global
block transforms the learnable query \( q \) into a video token \( X^G_i \). In this paper, we explore four fusion strategies to integrate the video tokens from all the global blocks \( \{X^G_i\}_{i=1} \) into a final video representation \( F \), and employ the sequential way to conduct fusion regarding efficacy and efficiency.

(a) **Sequential**: We sequentially use the video token from the previous global block \( X^G_{i-1} \) as the query token in the current global block \( q_i \), where \( X^G_i = G_i(X^G_{i-1}, X^L_i) \).

(b) **Parallel**: We concatenate all the tokens \( \{X^G_i\}_{i=1} \) in parallel, and use a linear projection \( U^F \in \mathbb{R}^{N \times C} \) to obtain the final token, where \( F = \text{Concat}(X^G_1, ..., X^G_N)U^F \).

(c) **Hierarchical KV**: We use the video token from the previous global block \( X^G_{i-1} \) as a part of contextual tokens in the current global block, where \( X^G_i = G_i(q_i, [X^G_{i-1}, X^L_i]) \).

(d) **Hierarchical Q**: We use the video token from the previous global block \( X^G_{i-1} \) as a part of query tokens in the current global block, i.e., \( X^G_i = G_i([X^G_{i-1}, q_i], X^L_i) \).

Finally, we extract the class token \( F^C \) from the final local UniBlock, and add it with the video token \( F \) by weighted sum, i.e., \( Z = \alpha F + (1-\alpha)F^C \), where \( \alpha \) is a learnable parameter processed by the Sigmoid function.

### 4. Experiments

**Datasets.** To evaluate the learning capability of our UniFormerV2, we conduct experiments on 8 popular video benchmarks, including the trimmed videos less than 10 seconds, and the untrimmed videos more than 1 min. The trimmed video benchmarks include: (a) Scene-related Kinetics, i.e., Kinetics-400, 600 and 700; (b) Heterogeneous Moments in Time V1 [44]; (c) Temporal-related Something-Something V1/V2 [22]. For the untrimmed video recognition, we choose ActivityNet [25] and HACS [78]. More dataset details can be found in supplemental materials.

**Kinetics-710 for Post-Pretraining** We propose a unified video benchmark for post-pretraining UniFormerV2. Different from [72] that exploits a web-scale video dataset (i.e., 60M video-text pairs), we build a much smaller video benchmark based on the Kinetics-400/600/700. Concretely, we merge the training set of these Kinetics datasets, and then delete the repeated videos based on Youtube IDs. Note that we have removed testing videos from different Kinetics datasets leaked in our combined training set for correctness. As a result, the total number of training videos is reduced from 1.14M to 0.66M. Additionally, we merge the action categories in these three datasets, which leads to 710 classes in total. Hence, we call this video benchmark Kinetics-710. In our experiments, we demonstrate the effectiveness of Kinetics-710. For post-pretraining, we simply use 8 input frames and adopt the same hyperparameters as training on the individual Kinetics dataset. After that, no matter how many frames are input (16, 32, or even 64), we only need 5-epoch finetuning for more than 1% top-1 accuracy improvement on Kinetics-400/600/700 (see Table 6).

**Implement Details.** Unless stated otherwise, we follow most of the training recipes in UniFormer [35], and the detailed training hyperparameters can be found in supplemental materials. We build UniFormerV2 based on ViTs pretrained with various supervisions (see Table 5), showing the generality of our design. For the best result, we adopt CLIP-ViT [50] as the backbone by default, due to its robust representation pretrained by vision-language contrastive learning. For most datasets, we insert the global UniBlocks in the last 4 layers of ViT-B/L to perform the multi-stage fusion. But for Sth-Sth V1/V2, we insert the global UniBlocks in the last 8/16 layers of ViT-B/L for better temporal modeling. The corresponding ablation studies are shown in Table 1, 2, 3. Finally, we adopt sparse sampling [63] with the resolution of 224 for all the datasets.
To evaluate the effectiveness of UniFormerV2, we investigate each key structure design. All the models are directly finetuned from CLIP-ViT-B/16 by default. We utilize “8×4×3” and “16×1×3” testing strategies for kinetics and Something-Something respectively.

**Different Components.** Table 1 indicates that the global UniBlock is crucial for the scene-related benchmark (e.g., K400), since it can effectively provide holistic video representation for classification. Alternatively, the local UniBlock is critical for the temporal-related benchmark (e.g., SthSthV2), as it can efficiently describe detailed video representation. Furthermore, using temporal downsampling with double input frames (similar FLOPs) enlarges the temporal receptive field, which is also helpful for distinguishing complex temporal-related actions.

**Local UniBlock.** To explore the structure of local UniBlock, we conduct experiments in Table 2. It reveals that convolution is superior to self-attention for temporal modeling, and our local MHRA outperforms both methods. Following ST-Adapter [47], we add another local MHRA after the spatial MHRA for better performance. To achieve the best accuracy-FLOPs trade-off, local MHRA is incorporated in all layers while reducing the channel by 1.5 times.

**Global UniBlock and Multi-stage Fusion.** Table 3 reveals that the features in the deep layers are critical for capturing long-term dependency, while the DPE and the middle information are necessary for identifying the motion difference. Furthermore, Table 4 shows that the simplest sequential fusion is adequate for integrating multi-stage features.

**Pretraining Sources.** To demonstrate the generality of our UniFormerV2 design, we apply it to the ViTs with various pertaining methods, including supervised learning [18, 56], contrastive learning [8, 50] and mask image modeling [24, 3]. Table 5 indicates that all the models beat TimeSformer [4], especially for SthSth V2 which relies on robust temporal modeling. The findings also suggest that a well-pretrained ViT enhances video performance.

**Training Recipes.** We compare different training and finetuning methods in Table 6. Note that when co-training with K400, K600 and K700, we remove the leaked videos in the validation set and introduce three classification heads. While K710 has only around 58% of the total training videos (0.66M vs. 1.14M for K400+K600+K700), it significantly enhances performances on Kinetics. Moreover, it decreases training costs by about 33%. Furthermore, direct training on K710 proves to be more effective than Kinetics co-training, especially for K600 (+1.3% vs. +1.0%) and K700 (+0.5 vs. -0.2%). Though co-finetuning shared the backbone and saved parameters, we individually fine-tune each dataset for better performance.

**Visualization.** In Figure 4, we compared UniFormerV2 with TimeSformer [4] and UniFormerV1 [35]. We use CAM [80] to show the most discriminative features that the
### Table 7: Results on scene-related Kinetics-400.

Our UniFormerV2 with public sources outperforms most of the current methods in terms of accuracy and/or efficiency. And it firstly achieves **90.0% top-1 accuracy** on Kinetics-400.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Pretraining Data</th>
<th>Frame × Crop × Clip</th>
<th>Param (M)</th>
<th>FLOPs (T)</th>
<th>K400 Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SlowFast101 [21]</td>
<td></td>
<td></td>
<td>80 × 3 × 10</td>
<td>60</td>
<td>7.0</td>
<td>81.8</td>
<td>95.1</td>
</tr>
<tr>
<td>MoViNet-A5 320 [35]</td>
<td></td>
<td></td>
<td>120 × 1 × 1</td>
<td>16</td>
<td>0.3</td>
<td>82.7</td>
<td>95.7</td>
</tr>
<tr>
<td>MTViTv1-B [19]</td>
<td>ViT-B</td>
<td></td>
<td>40 × 3 × 4</td>
<td>218</td>
<td>45.5</td>
<td>87.9</td>
<td>97.9</td>
</tr>
<tr>
<td>VideoMAE-B [54]</td>
<td>VT-B</td>
<td></td>
<td>16 × 3 × 5</td>
<td>87</td>
<td>2.7</td>
<td>81.5</td>
<td>95.1</td>
</tr>
<tr>
<td>TimeSformer-L [4]</td>
<td>VT-L</td>
<td></td>
<td>8 × 3 × 1</td>
<td>115</td>
<td>0.4</td>
<td>84.0</td>
<td>96.3</td>
</tr>
<tr>
<td>UniFormerV2-L [35]</td>
<td>VT-L</td>
<td>CLIP-400</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>87.3</td>
<td>97.7</td>
</tr>
<tr>
<td>UniFormerV2-L [35]</td>
<td>VT-L</td>
<td>CLIP-400</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>87.7</td>
<td>97.9</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>VT-L</td>
<td>CLIP-400+K710-0.66M</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>87.9</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>VT-L</td>
<td>CLIP-400+K710-0.66M</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>89.3</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>VT-L</td>
<td>CLIP-400+K710-0.66M</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>89.7</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>VT-L</td>
<td>CLIP-400+K710-0.66M</td>
<td>8 × 3 × 4</td>
<td>354</td>
<td>2.0</td>
<td>90.0</td>
<td>98.4</td>
</tr>
</tbody>
</table>

### Table 8: Results on scene-related Kinetics-600.

<table>
<thead>
<tr>
<th>Method</th>
<th>Frame × Crop × Clip</th>
<th>Param (M)</th>
<th>FLOPs (T)</th>
<th>K600 Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SlowFast101 [21]</td>
<td>80 × 3 × 10</td>
<td>60</td>
<td>7.0</td>
<td>81.8</td>
<td>95.1</td>
</tr>
<tr>
<td>MoViNet-A5 320 [35]</td>
<td>120 × 1 × 1</td>
<td>16</td>
<td>0.3</td>
<td>82.7</td>
<td>95.7</td>
</tr>
<tr>
<td>MTViTv2-L [37]</td>
<td>40 × 3 × 4</td>
<td>218</td>
<td>45.5</td>
<td>87.9</td>
<td>97.9</td>
</tr>
<tr>
<td>ClIP-CLIP-L [75]</td>
<td>16 × 3 × 4</td>
<td>430</td>
<td>7.9</td>
<td>89.3</td>
<td>97.7</td>
</tr>
<tr>
<td>CoVeR 448 [75]</td>
<td>16 × 3 × 1</td>
<td>431</td>
<td>17.6</td>
<td>87.9</td>
<td>97.7</td>
</tr>
<tr>
<td>CoCa 576 [73]</td>
<td>32 × 3 × 4</td>
<td>1000+</td>
<td>N/A</td>
<td>89.4</td>
<td>98.3</td>
</tr>
<tr>
<td>MTViH [72]</td>
<td>32 × 3 × 4</td>
<td>1000+</td>
<td>44.5</td>
<td>89.6</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L</td>
<td>32 × 3 × 2</td>
<td>354</td>
<td>16.0</td>
<td>89.5</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>32 × 3 × 2</td>
<td>354</td>
<td>16.0</td>
<td>89.5</td>
<td>98.3</td>
</tr>
<tr>
<td>UniFormerV2-L [33]</td>
<td>64 × 3 × 3</td>
<td>354</td>
<td>75.3</td>
<td>90.1</td>
<td>98.5</td>
</tr>
</tbody>
</table>

### 4.2. Comparison to state-of-the-art

**Kinetics.** Table 7 reports the results on scene-related Kinetics-400. (1) Compared with the advanced MViT2-L [37], which is specialized for video and requires prolonged image pertaining, our UniFormerV2-L achieves 1.2% higher performance with only 5% FLOPs. (2) Though VideoMAE [54] demonstrates that the vanilla ViT can be a strong video learner, it has to train the model from scratch for 1600 epochs, while our method effectively utilizes well-prepared ViTs to achieve significant improvement (87.3% vs. 85.2% with similar FLOPs). (3) The third part lists...
our counterparts based on image ViTs. Compared with the popular prompt tuning [47, 40], our method fully unlocks the potential of pretraining ViTs with remarkable improvement. For example, at similar FLOPs, our UniFormerV2-B achieves 1.1% and 2.0% higher top-1 accuracy than EVL-B [40] and ST-Adapter-B [47], respectively. Compared with X-CLIP-L [46] that utilizes the extra language knowledge, our UniFormerV2-L obtains 0.6% performance gain (87.7% vs. 87.1%). It is noteworthy that our single model, which only requires 1% video post-pretraining and 35% parameters, outperforms MTV-H [72] by achieving a new state-of-the-art result of 90.0% on Kinetics-400. As for Kinetics-600 and 700, our model also obtains the state-of-the-art performances (90.1% and 82.7%, see Table 8 and 9).

**Moments in Time.** Due to complex inter-class and intra-class variations, MiT is more challenging than Kinetics. As shown in Table 10, our model beats most of the recent methods, e.g., compared with ViViT-L [1], UniFormerV2-B obtains 4.2% performance gain but only with 19% model parameters and 15% FLOPs. Compared with MTV-H [72], UniFormerV2-L only uses 35% model parameters and 25% FLOPs to achieve 2.2% top-1 accuracy improvement.

**Something-Something.** Table 11 presents the results on temporal-related SthSth V2. It reveals that the existing state-of-the-art methods are specialized or based on masked modeling, both of which require expensive pretraining. In contrast, our method is economically friendly, as it uses open-source ViTs. UniFormerV2-L achieves comparable performance with the latest MTVv2-L [37] (top-1: 73.0% vs. 74.3% and VideoMAE-L [54] (top-5: 94.5% vs. 94.6%). Furthermore, the results demonstrate that previous plug-and-play methods perform much worse on the temporal-related task. For example, EVL-L [40] achieves 1.1% higher performance than VideoMAE-L on K400, but obtains 7.6% lower accuracy on SthSthV2. However, our method can arm image ViT for strong temporal modeling, delivering 6.4% performance gain than EVL with fewer computation costs on SthSth V2. Additionally, for Sthsh V1 in Table 12, we achieve the new state-of-the-art performance (62.7%). These results demonstrate the effectiveness and efficiency of UniFormerV2 for temporal modeling.

**ActivityNet and HACS.** For the untrimmed videos, it is essential to capture long-range temporal information, since the action may occur multiple times at arbitrary moments. As shown in Table 13, our UniFormerV2 significantly outperforms the previous best methods on the large-scale untrimmed benchmark ActivityNet and HACS by 4.5% and 3.6%, respectively. These results demonstrate the strong long-term modeling capacity of our method.
5. Conclusion

In this paper, we serve UniFormer as efficient plug-and-play modules for image ViTs, enhancing their abilities as strong video learners. Extensive experiments demonstrate that our UniFormerV2 can unlock the full potentials of image ViTs, achieving state-of-the-art performances on 8 large-scale benchmarks. To the best of our knowledge, it is the first model to reach 90% top-1 accuracy on Kinetics-400. As the research community becomes increasingly open, we hope our method will be instrumental in building powerful yet cost-effective video foundation models.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160100, No. 2022ZD0160505, No. 2022ZD0160900), the National Natural Science Foundation of China (No. 62076119), the Joint Lab of CAS-HK, the National Natural Science Foundation of China under Grant (No. 62272450), the Shenzhen Research Program (RCJC20200714114557087), and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020355).

References

[14] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? ArXiv, abs/2102.10882, 2021. 2
[19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In IEEE/CVF International Conference on Computer Vision, 2021. 2, 3, 7, 8
[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and...


[61] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding, 2023. 2


[75] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M. Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *ArXiv*, abs/2112.07175, 2021. 7, 8

[76] Chen-Lin Zhang, Jian Zhai Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *ArXiv*, abs/2202.07925, 2022. 2


