Integrating Pose and Mask Predictions for Multi-person in Videos

Miran Heo[1*] Sukjun Hwang[1] Seoung Wug Oh[2] Joon-Young Lee[2] Seon Joo Kim[1]
[1]Yonsei University [2]Adobe Research
https://miranheo.github.io/human-vis

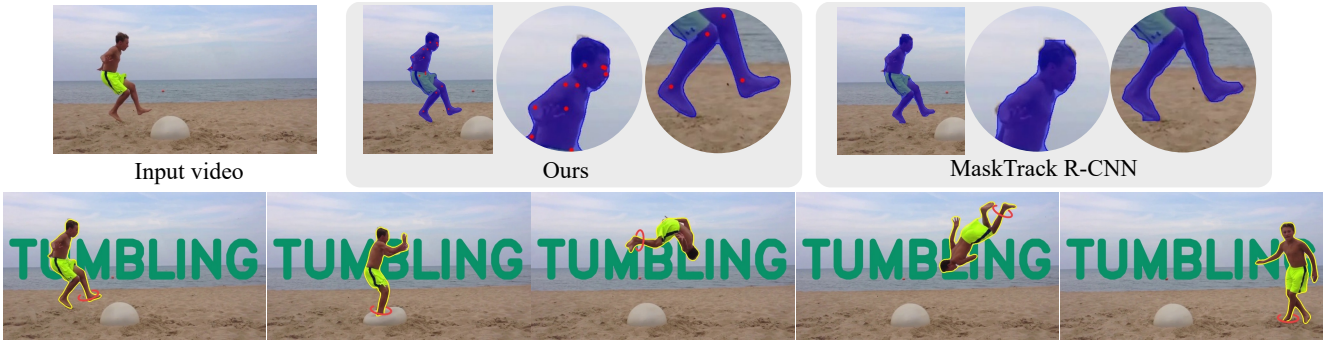

Figure 1. Our model simultaneously tracks, finds body joints, and segments high-quality mask for all human instances in a video. Compared to off-the-shelf online video instance segmentation models, our method generates masks of finer granularity (first row). Also, jointly tracking pose and mask allows us to apply multiple visual effects because segmentation masks help layer separation while joint information supports motion tracking (second row).

Abstract

In real-world applications for video editing, humans are arguably the most important objects. When editing videos of humans, the efficient tracking of fine-grained masks and body joints is the fundamental requirement. In this paper, we propose a simple and efficient system for jointly tracking pose and segmenting high-quality masks for all humans in the video. We design a pipeline that globally tracks pose and locally segments fine-grained masks. Specifically, CenterTrack is first employed to track human poses by viewing the whole scene, and then the proposed local segmentation network leverages the pose information as a powerful query to carry out high-quality segmentation. Furthermore, we adopt a highly light-weight MLP-Mixer layer within the segmentation network that can efficiently propagate the query pose throughout the region of interest with minimal overhead. For the evaluation, we collect a new benchmark called KineMask which includes various appearances and actions. The experimental results demonstrate that our method has superior fine-grained segmentation performance. Moreover, it runs at 33 fps, achieving a great balance of speed and accuracy compared to the prevailing online Video Instance Segmentation methods.

1. Introduction

With explosive demand for video contents and advances in computer vision technology, deep learning-based video editing techniques have been successfully used in professional desktop programs and mobile applications for non-professional users. For example, algorithms for multi-person pose tracking [1, 20] can be applied to a variety of applications where human joint information is necessary such as animation and movie editing. In addition, video object segmentation [29] allows users to obtain sophisticated masks for the objects guided by an initial mask or scribbles, enabling various masking-based visual effects such as composition.

Recently, the task of video instance segmentation (VIS) [44] has been introduced, which predicts masks of objects in videos without user-provided guidance. However, we argue that VIS algorithms are usually not optimal for video editing applications for following reasons. First, most VIS methods are optimized for YouTube-VIS benchmark [44] that is composed of categories that are less likely to be used by editors. In addition, the metric used for the evaluation (AP) makes existing approaches to value detection and tracking over segmentation qualities. As a result, the VIS task becomes impractical in the perspective of editors, whom frequently demand human masks of fine-granularity with high efficiency.

*Work done during an internship at Adobe Research.

When applying visual effects on humans, the pose information is highly required in addition to the mask. Fine grained segmentation masks and human-specific articular information are complementary to each other. As shown in Fig. 1, segmentation masks help layer separation while joint information supports local patch tracking and those information are the basic needs for the end-users. While there exists a notable number of works that tackle each problem separately, jointly tracking poses and segmenting masks for human instances has not been thoroughly addressed in the literature.

In this paper, we propose a simple and efficient online video human instance segmentation and pose tracking pipeline which is applicable to daily videos. As shown in the illustration in Fig. 2, we take a *globally track pose then locally segment mask* strategy. Our intuition is that lightweight but accurate segmentation module can be designed using human-specific skeleton information as a powerful query. Specifically, for Global Pose Tracker, we adopt CenterTrack [47] which was initially proposed for multi-object tracking but also demonstrates excellent flexibility in human pose estimation. CenterTrack efficiently detects and tracks all the human instances in the video. To have a large receptive field that cover the whole scene and make the inference of tracker faster, we down-scale the input from the original resolution.

Next, segmentation is performed locally for each detected human instance by cropping the region of interests (ROIs) from the original input. We maintain high resolutions for each ROIs to keep fine-details. Each cropped ROIs are concatenated with their corresponding joint heatmap, then passed to the segmentation module. Note that the heatmap contains explicit pose information about a person. Therefore, our goal is to design a highly efficient segmentation network that effectively understands the context of the ROI by utilizing the heatmap and mostly dedicate to focus on the details. From experiments, shallow networks with typical CNN blocks (*e.g.* ResNet-18) inherently show limitations encoding sparsely distributed heatmaps due to small receptive fields. In order to improve the mask quality, an off-the-shelf solution would be to stack more CNN blocks or use extra modules such as ASPP [11, 12] to enlarge the receptive field. However, adopting such architectures inevitably leads to more computations, resulting in low efficiency.

To overcome the issue, we adopt recently proposed MLP-Mixer (Mixer) [33] as an efficient global information gatherer. Since its architecture is based entirely on multilayer perceptrons (MLPs), the additional overhead is extremely low while taking the advantage of the global receptive field within ROI. We insert a single Mixer layer to the encoder of our segmentation module with the motivation of effectively propagating the context which is implicitly encoded in the heatmaps to the overall feature at once. Finally, our model remains highly efficient while preserving fine details and segmenting masks of high quality. It is noteworthy that this is the first study to use Mixer in combination with CNN to increase the efficiency of the dense predication task.

The current VIS dataset is not human-centric and the ground truth of the validation dataset is not provided with only full class score available. Therefore, we introduce a new benchmark called KineMask for the evaluation which focuses on human instances with diverse appearances and motions. To demonstrate the effectiveness of our method, we also propose a new metric that measures the accuracy of boundary region along the temporal axis. The proposed approach is very simple and runs at 33 FPS on a single RTX 2080Ti GPU, while achieving strong results.

Our main contributions are summarized as follows:

- We tackle the problem of jointly tracking pose and segmenting masks for all humans in a video for the first time. This problem has been overlooked in the computer vision community.
- To design an efficient architecture, we propose a pipeline which globally tracks pose then locally segments mask, and exploit human pose information as a strong query to segment the mask.
- To address the small receptive field issue of the shallow network, we adopt Mixer layer to efficiently aggregate local features.
- We introduce a human-centric video instance segmentation benchmark and the metric for measuring the accuracy of boundary regions in a video.

2. Related Work

2.1. Video Instance Segmentation

VIS aims to simultaneously predict categories, segmentation masks, and identities of all instances in a video. Due to the complexity of the problem, prevailing approaches follow a track-by-detect paradigm that associates instances throughout a video after performing frame-level instance segmentation [9, 25, 44, 45]. On the other hand, several methods exploit clip-level information and demonstrate a robustness on occlusion or motion blur [2, 5, 19, 39].

Online methods. MaskTrack R-CNN [44] designs additional tracking branch on the strong two-stage instance segmentation baseline, Mask R-CNN [16]. They assign identities by computing the criterion which includes the pairwise cosine similarity of visual feature obtained from the tracking branch and extra traditional matching logics (e.g.,

bounding box IoUs and class id). On the other hand, SipMask [9] and SGNet [25] extend single-stage detector YOLACT [7] and FCOS [32] respectively. SipMask proposes a light-weight spatial preservation module that preserves the spatial information within a bounding box and generates separate set of spatial coefficients for each bounding box sub-region, enabling improved delineation of spatially adjacent objects. SGNet builds additional mask head and tracking head on FCOS detector. Thanks to the fully convolutional nature of FCOS, their mask prediction dynamically performs spatial attention on each sub-region of instance that leads to a fine mask quality than RoI based methods [8, 16]. While aforementioned methods follow the top-down approach, STEm-Seg [2] takes a clip input and considers a video as 3D spatio-temporal volumes and learns to segment instances in videos in a bottom-up fashion by leveraging spatio-temporal embeddings. CrossVIS [45] correlates feature spaces of different frames by applying dynamically generated filters to different frames.

Offline methods. Given a short video clip, MaskProp [5] propagates instance-specific features in the center frame obtained from Mask R-CNN [16] to its neighboring frames. It alleviates challenging problems such as occlusion and motion blur by utilizing clips in a video as windows and performing clip-level matching. VisTR [39] extends a transformer based detector DETR [10] to the VIS task. By taking the entire video as input and processing it at once, VisTR solves the VIS from a new perspective of similarity learning.

2.2. Multi-Object Tracking and Segmentation

Multi-Object Tracking and Segmentation (MOTS) extends the popular task of multi-object tracking task (MOT) to instance segmentation. Similar to MaskTrack R-CNN, [35] suggests a baseline named Track R-CNN which applies extra tracking branch built upon Mask R-CNN. PointTrack [43] generates a new tracking-by-points paradigm where discriminative instance embeddings are learned from randomly selected points rather than images. TraDes [40] presents an online joint detection tracking methods based on CenterNet [48] and infers object as an offset by a cost volume. MOTS is composed of videos with numerous pedestrians, thus its target applications are autonomous driving, video surveillance, and robotics. In this paper, however, we do not directly aim for such scenarios.

2.3. Multi-Person Pose Tracking

Understanding human objects has long been an important issue in the field of computer vision, as it is widely applied to many fields such as human interactions, action recognition, video surveillance, and sports video analysis. With the emergence of large-scale benchmark datasets [1, 20], numerous approaches have addressed the problem of articulated multi-person body joint tracking in monocular video, and made significant progress [15, 21, 30, 31, 37, 41, 42]. Despite of recent advances in human pose tracking, human-specific instance segmentation in videos has not been thoroughly addressed in the literature.

2.4. Global Receptive Field Models

From the nature of convolutional filters, attending only local information, the computer vision community have been studying the effects of receptive fields. With the emergence and massive studies on deep learning, it is proven effective to gradually increase the receptive field by stacking multiple convolutional layers, and employ multiple features of different receptive fields as SIFT [27]. In particular, ASPP [12] module dramatically increases the receptive field by associating atrous convolutions and global average pool, and adopting the module leads to improvement of performance on various vision tasks.

As an alternative to ASPP, obtaining the global receptive field [34, 38] using attention [4] based methods is getting spotlighted. Taking the global information into account, an use of the methods leads to a great accuracy. However, the attentional approaches suffer from quadratic increase of computation with respect to the number of input. Therefore, many researchers are focusing on diminishing the overhead while remaining powerful, *i.e.*, decompsing the attention [6, 36].

Recently, Mixer [33] have shown that the attention is indeed not the necessary for the global receptive field. The simplicity of Mixer highly reduces the overall computation, thus highly efficient. Our model utilizes a Mixer layer for understanding semantic information in cropped patches of a high resolution. The joint heatmap of a person gives a strong clue to the pose of the targeted person. Without heavy computations derived from attentional methods, the Mixer layer effectively encodes the pose information with its global receptive field.

3. Method

We propose a *globally track pose then locally segment mask* strategy for the final goal of designing an efficient system that simultaneously predicts track IDs and generates masks with fine granularity for all humans over the entire video. Our intuition is that human-specific skeleton information can be a powerful query for finding foreground masks rather than bounding boxes, thus allowing us to devise a light-weighted segmentation model while generating masks of high accuracy. The overview of our method is shown in Fig. 2.

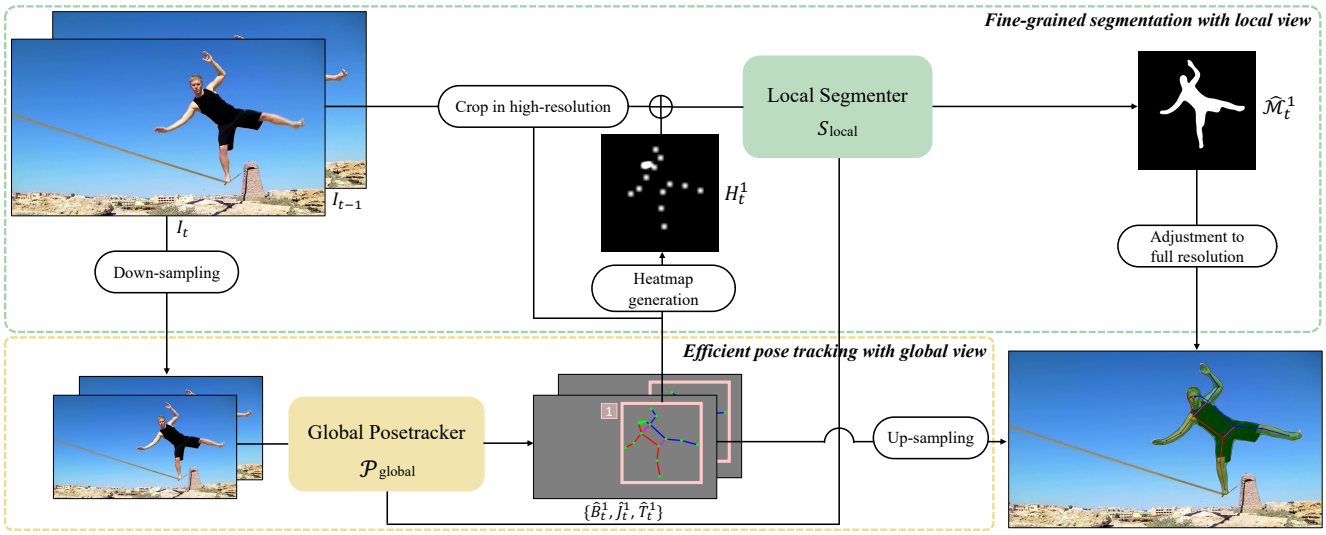

Figure 2. Overview of our framework. Our framework consists of two modules: $\mathcal{P}_{\text{global}}$ and $\mathcal{S}_{\text{local}}$. Given two consecutive frames, $\mathcal{P}_{\text{global}}$ globally tracks human body joints for all humans in the current frame. Then, $\mathcal{S}_{\text{local}}$ takes cropped RGB images with additional joint agnostic heatmap and results fine grained segmentation masks.

3.1. Global Pose Tracker

Our method is built on the CenterTrack [47] as a tracker with an additional pose estimation head introduced in [48]. At time t, given the current frame I_t and the prior frame I_{t-1} with an instance-agnostic centerness heatmap C_{t-1} which represents tracked objects in the previous frame, CenterTrack predicts the bounding boxes, joint coordinates, and track ids for every human in the current frame.

$$\{\hat{B}_t^i, \hat{J}_t^i, \hat{T}_t^i\}_{i=1}^{N_t} = \mathcal{P}_{\text{global}}(I_t, I_{t-1}, C_{t-1}), \tag{1}$$

where $\hat{B}_t^i$ is the bounding box, $\hat{J}_t^i$ is the k 2D human joint locations (k is 17 in COCO [24]), and $\hat{T}_t^i$ is the track id of the i^{th} object. N_t is the number of detected instances in the current frame. To have a large receptive field that cover the whole scene and make the inference of tracker faster, the tracker network takes down-sampled frames (512×512). Note that our method possesses good modularity (not dependent on CenterTrack), thus Global Pose Tracker can be replaced by any other pose tracking architecture depending on the purpose.

3.2. Local Segmenter

To increase efficiency of the system, relatively less detail-critical elements are processed globally, and segmentation masks that require pixel-level details are concentrated and processed locally. The goal of Local Segmenter $\mathcal{S}_{\text{local}}$ is that finding fine grained masks for all humans in the scene at time t.

Local region cropping. Prevailing off-the-shelf two-stage detectors suffer from low segmentation quality driven from the coarse resolution of 28×28 [16]. In order to preserve fine details and bring high quality, we take a cropped region of a high resolution from the original RGB frame as an input to the Local Segmenter. Our Local Segmenter focuses on the specific sub-region, generating much fine-grained masks than previous two-stage detectors. We take predicted bounding boxes $\hat{B}_t$ from $\mathcal{P}_{\text{global}}$ and expand it by 1.5 times to relax the tight bounding boxes, and up-sample the cropped region to 512×512 as input to the network.

Joint heatmap generation. We exploit body joint coordinates from $\mathcal{P}_{\text{global}}$ as a strong query to generate human instance masks. The coordinates of predicted joint points are re-adjusted for the corresponding box area. After that, we create a single joint-agnostic heatmap with activations in the regions of the points so that the joints information for the target instance are given as a guiding signal to the input of the local segmentation module. We center a 2D Gaussian around each of the points, in order to create a single heatmap. The heatmap is concatenated with the RGB channels of the cropped input image, to form a 4-channel input for the local segmentation module. Then, the final results at time t frame are computed as follows:

$$\{\hat{\mathcal{M}}_t^i\}_{i=1}^{N_t} = \mathcal{S}_{\text{local}}(\{\hat{B}_t^i(I_t) \oplus G_t^i\}_{i=1}^{N_t}), \tag{2}$$

where $\hat{\mathcal{M}}$ is predicted masks, G is joint heatmap, and $\oplus$ denotes matrix concatenation.

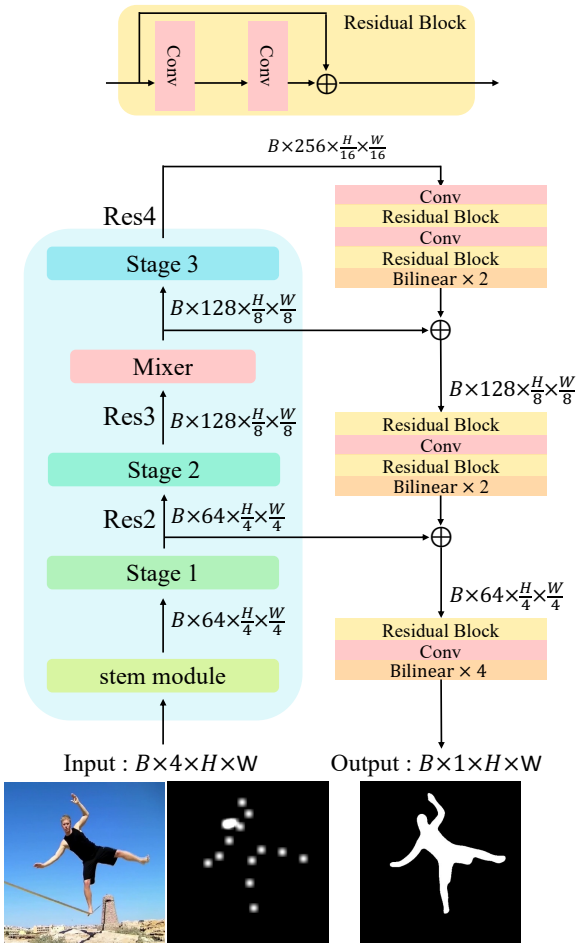

Figure 3. Detailed architecture of Local Segmenter.

Mixer layer. Despite the heatmap with explicit joint location provides stronger foreground information than bounding boxes, a way to effectively transfer it to the network is needed because each point is located sparsely. A natural approach will be using a deep encoder-decoder structure with large receptive fields such as ResNet-50 [17] with ASPP [11]. However, adopting such an architecture inevitably consumes expensive computations, which is less efficient.

To retain large receptive fields while maintaining lightweight networks, we leverage recently introduced Mixer [33]. Mixer is a simple architecture which is entirely composed of MLPs and attains competitive result on image classification task as convolutions or self-attention. As the name indicates, the core idea of Mixer is that repeatedly mixing the features across either spatial locations or feature channels allows global communication.

Based on our empirical observation that shallow networks struggle to associate sparsely located joint information, we employ a single Mixer layer after several CNN blocks propagates localized visual features. Therefore, our network enjoys both global association of sparsely annotated joint heat-points and the rich localization information of foreground objects while maintaining efficiency.

Specifically, we insert single Mixer layer on Res3 features of $H'W'$ ($1/8$ to the original resolution) and the output feature $\mathbf{Y}$ is computed as follows:

$$\begin{aligned}\mathbf{U}_{*,i} &= \mathbf{F}_{*,i} + \mathbf{W}_2\sigma(\mathbf{W}_1 LN(\mathbf{F})_{*,i}), \text{ for } i = 1...c, \\ \mathbf{Y}_{j,*} &= \mathbf{U}_{j,*} + \mathbf{W}_4\sigma(\mathbf{W}_3 LN(\mathbf{U})_{j,*}), \text{ for } j = 1...s,\end{aligned} \tag{3}$$

where $\mathbf{F}$ is Res3 features, c is hidden dimension, and $s = H'W'/p^2$ is the number of patches. Here LN indicates Layer Normalization [3] and σ is an activate function (GELUs [18]). We use $c = 128$ and $p = 1$. It is noteworthy that this is the first study to use Mixer in combination with CNN to increase the efficiency of the dense predication task.

Architecture. Fig. 3 illustrates details of our local segmentation network. We adopt encoder-decoder structure with skip-connections. For encoder network, we use Res2, Res3, Res4 feature of ResNet-18. And single Mixer layer aggregates global information adopting Res3 feature. Decoder network consists of conventional residual blocks to refine the final mask. The overall loss for the Local Segmenter is written as:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{dice}, \tag{4}$$

where $\mathcal{L}_{bce}$ is Binary cross entropy loss and $\mathcal{L}_{dice}$ is Dice loss [28].

4. Training details

Global Pose Tracker. We adopt CenterTrack with pose estimation head and DLA34 [46] is used as a backbone. The module is trained using COCO Person dataset [24] and optimized with AdamW [26] with learning rate 1e-4.

Local Segmenter. Whereas to inference where only predicted bounding boxes are used for the sub-regions, we improve the robustness of our model and stabilize training by alternately assigning the sub-regions with ground truth boxes B^* and predicted bounding boxes $\hat{B}$. Specifically, the assignment is decided by intersecting areas between B^* and $\hat{B}$. Given M ground truths and N box predictions, the matching score $\mathcal{O} \in [0,1]^{M\times N}$ is calculated as follows:

$$\mathcal{O}(m,n) = \frac{Area(B^*_m \bigcap \hat{B}_n)}{Area(B^*_m)}, \tag{5}$$

where *Area* denotes the area of the given box and $\bigcap$ results an intersecting box of the given box pair. Then, we pair

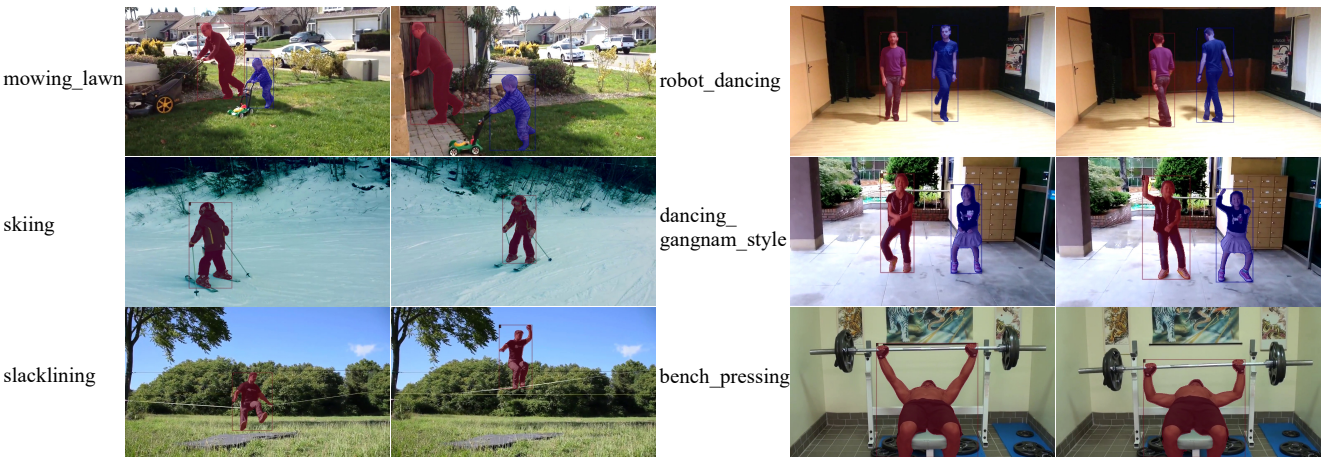

Figure 4. Examples of KineMask benchmark.

up each predicted box with a ground truth box as $\hat{\sigma}(n) = \arg\max_m \mathcal{O}(m, n)$. The assignments of the sub-regions r are finalized as follows:

$$r_i = \begin{cases} \hat{B}_i & \text{if} \quad \mathcal{O}(\hat{\sigma}(i), i) \geq \tau \\ B^*_{\hat{\sigma}(i)} & \text{if} \quad \mathcal{O}(\hat{\sigma}(i), i) < \tau \end{cases}, \tag{6}$$

which alternates between a predicted box and a ground truth box by the threshold τ. If the intersecting area of a predicted box is above τ, the predicted box is assigned as it covers the majority of the corresponding ground truth box. On the other hand, if the intersecting area is below τ, it is likely that the predicted box cannot sufficiently cover the mask within the ground truth box, thus the ground truth box is taken. The joint points are also assigned using the same logic as Eq. 6. In this paper, we use $\tau = 0.5$ after assigning optimal matching by Hungarian algorithm [23].

Local Segmenter is also trained using COCO Person dataset [24] and optimized with SAM [14] with learning rate 1e-4. We apply dropout to Mixer layer with probability 0.1.

5. Experiments

5.1. KineMask Benchmark

Currently available dataset for VIS is YouTube-VIS [44] and it contains 2,883 and 3,859 high-resolution YouTube videos for 2019 and 2021, respectively. Although it covers 40 categories including *person*, the ground truth of validation dataset is not available and only full class evaluation score is provided by the evaluation server, thus we cannot conduct experiments on only person class.

To this end, we collect a benchmark called KineMask for human video instance segmentation which consists of YouTube videos filmed with a variety of camera devices.

videos	frames	instances	action classes	resolution
500	19,154	720	79	720×1280

Table 1. Statistics of KineMask benchmark.

Data source. Kinetics dataset [22] is originally introduced for human action recognition and popularly used for training neural network architectures for understanding human behavior. For the source of our evaluation dataset, we adopt Kinetics-400; a large-scale video dataset, covering a diverse range of human actions.

Statistics. As summarized in the Table 1, KineMask consists of 500 high-resolution videos and has a resolution of 720×1280. The fps of video is generally 30 and we annotate all the masks at 6 fps. Also, the average video length is about 6 seconds. As our benchmark is collected from Kinetics, it includes 79 action classes, thus represents diverse appearance of objects and large range of motions. Fig. 4 illustrates some examples of KineMask benchmark.

5.2. Evaluation metrics.

Temporal Mask IoU. We follow the standard evaluation metric of VIS, proposed in [44]. The metric is based on the average precision (AP) and average recall (AR) which are standard evaluation metrics in image instance segmentation, and is extended to the temporal mask sequence. Given two mask sequences G and P, TMIoU at time t is formulated as follows:

$$\text{TMIoU}(G, P) = \frac{\sum_{t=1}^{T} |G_t \cap P_t|}{\sum_{t=1}^{T} |G_t \cup P_t|} \tag{7}$$

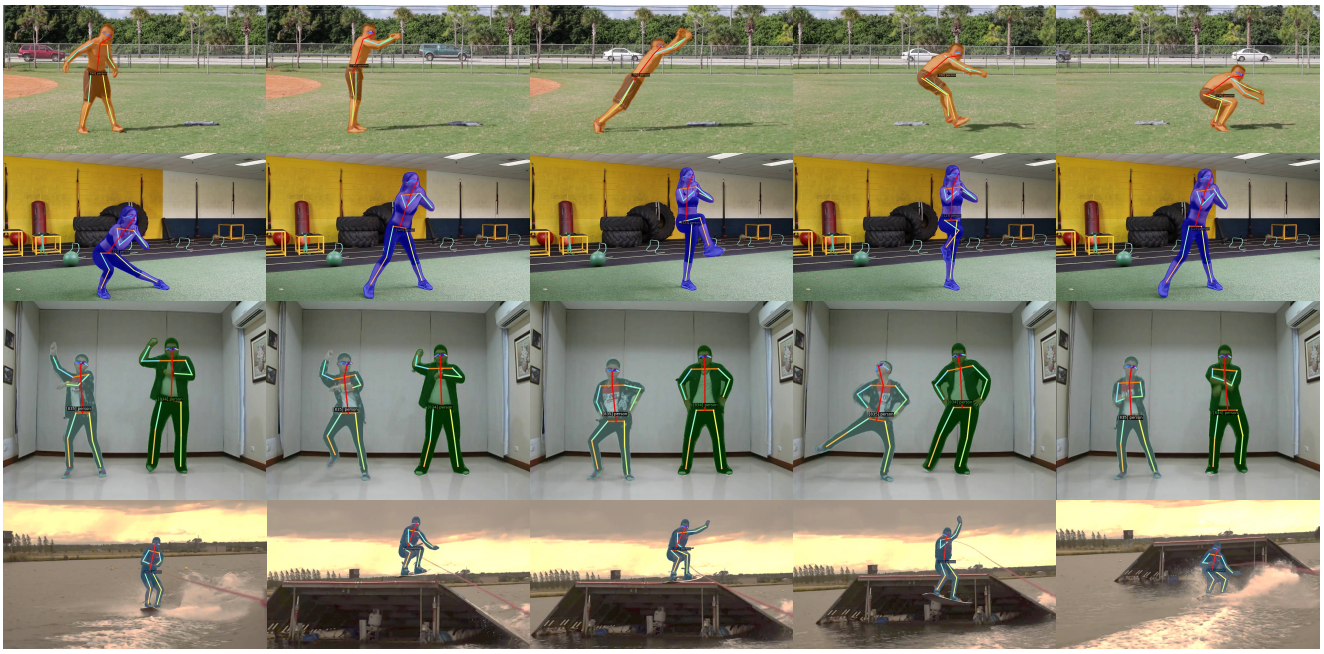

Figure 5. Visualization results. The mask color indicates unique instances.

2*Method	Training dataset		Prediction		2*FPS	Temporal Mask AP				Temporal Boundary AP		
	COCO	YTVIS	Mask	Keypoint		AP	AP_{85}	AP_{90}	AP_{95}	AP	AP_{50}	AP_{75}
MaskTrack R-CNN [44]†	✓	✓	✓		25.30	71.4	71.1	34.9	1.6	23.6	76.3	3.1
MaskTrack R-CNN [44]*	✓		✓		27.33	80.3	85.0	54.7	0.5	30.1	86.9	6.8
SipMask [9]†	✓	✓	✓		37.06	79.3	79.8	56.6	10.6	42.1	87.6	34.5
SipMask [9]*	✓		✓		38.65	81.4	83.7	64.2	10.8	40.9	89.8	27.2
Ours	✓		✓	✓	33.59	**84.2**	**88.0**	**76.3**	**22.3**	**45.9**	**91.5**	**42.8**

† reproduced using provided checkpoint

* retrained on COCO person class

Table 2. Comparison of our approach with the online VIS methods.

Temporal Boundary IoU. To evaluate both spatial-temporal consistency and mask quality of boundary region of predicted and ground truth masks, we propose new metric named Temporal Boundary IoU (TBIoU). Similar to YouTube-VIS metric, we extend image based Boundary IoU metric [13] to temporal axis. TBIoU is formulated as follows:

$$\text{TBIoU}(G, P) = \frac{\sum_{t=1}^{T} \left| (G_t^d \cap G_t) \cap (P_t^d \cap P_t) \right|}{\sum_{t=1}^{T} \left| (G_t^d \cap G_t) \cup (P_t^d \cap P_t) \right|}, \quad (8)$$

where boundary regions G_t^d and P_t^d are the sets of all pixels within d pixels distance at time t from the ground truth and prediction contours respectively. The dilation is computed by *dilation ratio*$\times$*image diagonal*. In this paper, we use 0.01 as ratio to calculate dilation.

5.3. Comparison with online VIS models

In this section, we compare proposed method against two online VIS models, MaskTrack R-CNN [44] and SipMask [9] in Table 2. The comparison is measured on the same machine, using a single RTX 2080Ti GPU. We used the official codes and checkpoints provided by the authors for the measurements. Since their original models are trained on 40 categories, we also provide results of models trained on COCO Person dataset. And we applied same schedule as the original repository.

For the temporal mask AP (TMIoU-measure), our method achieves 84.2 AP which outperforms all other published online VIS methods. More notably, for the temporal boundary AP (TBIoU-measure), our method achieves 45.9 AP which outperforms both MaskTrack R-CNN and SipMask with the stronger backbone ResNet-50. Moreover, the result of TBIoU-measure proves that our model has superior

Input Editing example

Figure 6. Application example of proposed method. As our method successfully generates both (1) segmentation masks and (2) human body joints, two different types of video editing can be applied simultaneously through entire video: (1) applying background effect for each person, and (2) the effect can be located specific body parts of the corresponding instance through multiple frames.

boundary accuracy compared to prevailing VIS approaches. Our model is highly efficient and runs at 33 fps which is competitive performance considering our method predicts both mask and keypoints. Overall performance shows that our method accomplishes strong quality and speed balance.

5.4. Ablation studies

We provide ablation studies and discuss how different settings affect overall performance of our method. The ablation studies are conducted on the COCO Person training set. Results are shown in Table 3 and discussed in detail next.

Keypoint information. We first investigate the effectiveness of additionally providing keypoint information which is represented as a heatmap to Local Segmenter. By providing more detailed localized information than the bounding box, the keypoint guidance brings a critical performance improvement. Given the keypoint information, both TMIoU and TBIoU AP scores increase significantly by +5.7% and +6.2% AP respectively.

Mixer layer. After the localization is conducted from Global Pose Tracker, the effectiveness of the Mixer layer is noticeable in the TBIoU-measure. While TMIoU AP improves marginally when adding the Mixer layer (+0.9%), TBIoU AP score highly improves by 2.0%. The result indicates that the Mixer layer improves the boundary details by aggregating sparse body joint information.

5.5. Qualitative Results and Applications

KineMask Benchmark. In Fig. 5, we show the results of our methods visualized on KineMask benchmark. We apply different colors to represent different instances. The visualized results suggest that our model generates segmentation masks of finer granularity under various action classes. Video results are included in our project page.

Settings		Temporal Mask AP			Temporal Boundary AP		
Keypoint	Mixer	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
		77.6	95.1	87.1	37.5	81.5	30.5
✓		83.3	95.3	91.8	43.7	89.1	37.6
✓	✓	84.2	96.0	92.7	45.7	91.5	42.8

Table 3. Ablations on the COCO Person training set.

Applications. We also provide a video editing example using our proposed method in Fig. 6. As our model predicts both human joint coordinates and high-quality segmentation masks, they can play different roles in real editing scenarios. First, binary masks for each instance can be used to separate layers so that we can apply effects such as changing backgrounds or moving foreground object. In this example, we apply three different background effects to different instances. Second, tracking keypoints over video can act as human-specific motion tracking. Thanks to the tracked coordinates of body joints, we exploit certain keypoint as an anchor point for the background effect. Therefore, our method provides two complementary information at the same time, while even being fast. More application examples are included in our project page.

6. Conclusion

We addressed the problem of jointly tracking pose and segmenting masks for all humans in the video, which has been overlooked in computer vision community. Our experiments demonstrated that proposed method produces high-quality masks with fast inference time while it is conceptually simple. Moreover, by presenting two complementary information simultaneously, our model exhibits high utilization for real-world video editing applications.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 1, 3

[2] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastianan Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 2, 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[5] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 2, 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 3

[7] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 3

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 3

[9] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020. 2, 3, 7

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[11] Liang-Cheih Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *IEEE TPAMI*, 2017. 2, 5

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 3

[13] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 7

[14] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *ICLR*, 2020. 6

[15] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *CVPR*, 2018. 3

[16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3, 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[18] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[19] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *arXiv preprint arXiv:2106.03299*, 2021. 2

[20] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *CVPR*, 2017. 1, 3

[21] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *CVPR*, 2019. 3

[22] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[23] Harold W Kuhn. The hungarian method for the assignment problem. In *Naval research logistics quarterly*, 1955. 6

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4, 5, 6

[25] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*, 2021. 2, 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. 2004. 3

[28] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[29] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1

[30] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *CVPR*, 2019. 3

[31] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In *CVPR*, 2020. 3

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 3

[33] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 2, 3, 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[35] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019. 3

[36] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 3

[37] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *CVPR*, pages 11088–11096, 2020. 3

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[39] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2020. 2, 3

[40] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. *arXiv preprint arXiv:2103.08808*, 2021. 3

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 3

[42] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018. 3

[43] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Huan Huang, Shilei Wen, Errui Ding, and Liusheng Huang. Segment as points for efficient online multi-object tracking and segmentation. In *ECCV*, 2020. 3

[44] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 6, 7

[45] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. *arXiv preprint arXiv:2104.05970*, 2021. 2, 3

[46] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, 2018. 5

[47] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 2, 4

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3, 4