Boosting Unsupervised Video Instance Segmentation with Automatic Quality-Guided Self-Training

Supplementary Material

6. Detailed methodology

This section supplements what is not clearly stated in Sec. 3.2.

6.1. Automated pseudo-annotation with spatiotemporal NMS

After the training of the VIS model, we use it to label the unlabeled videos. Let $\mathcal{D} = \{d_i\}_{i=1}^{N}$ denote the initial detection set per video, where each detection $d = (s_i, \{m_i^t\}_{t=1}^{T})$ contains:

- $s_i \in [0, 1]$: Confidence score
- $\{m_i^t\}_{t=1}^{T}$: Binary mask sequence across T frames

We filter those detections that have confidence scores larger than or equal to 0.25:

$$\mathcal{D}_{\text{filtered}} = \{d_i \in \mathcal{D} \mid s_i \geq 0.25\} \tag{16}$$

However, the detection sets may contain duplicate detections. To solve this problem, we need to perform spatiotemporal non-maximum suppression. First, we sort the detections based on their confidence scores:

$$\mathcal{D}_{\text{sorted}} = \{d_{(k)}\}_{k=1}^{|\mathcal{D}_{\text{filtered}}|} \quad \text{s.t.} \quad \forall i < j : s_{(i)} \geq s_{(j)} \tag{17}$$

Our method eliminates redundant detections through spatiotemporal overlap analysis: Any lower-confidence prediction is suppressed if exhibiting mask overlap (IoU $\geq$ 0.5) with higher-confidence detections in at least one video frame. Formally, let $\mathcal{D}_{\text{suppressed}} \subseteq \mathcal{D}_{\text{sorted}}$ represent the preserved detection set after suppression:

$$d_{(k)} \in \mathcal{D}_{\text{suppressed}} \iff \nexists d_{(p)} \in \mathcal{D}_{\text{suppressed}} \text{ where } p < k \text{ s.t. } \underbrace{\exists t \in [1, T] : \frac{\|m_{(k)}^t \cap m_{(p)}^t\|}{\|m_{(k)}^t \cup m_{(p)}^t\|} \geq 0.5}_{\text{Frame-specific overlap condition}} \tag{18}$$

where $\|\cdot\|$ denotes pixel cardinality. This temporal-existential criterion suppresses duplicates appearing in any frame of the video sequence.

6.2. Confidence-aware filtration via quality predictor

Let $\mathcal{D}_{\text{global}}^{(k)}$ denote the union of $\mathcal{D}_{\text{suppressed}}$ of all videos:

$$\mathcal{D}_{\text{global}}^{(k)} = \bigcup_{v \in \mathcal{V}} \mathcal{D}_{\text{suppressed},v}^{(k)} \tag{19}$$

where $\mathcal{D}_{\text{suppressed},v}^{(k)}$ denotes preserved detections for video v after spatiotemporal NMS at iteration k.

For each detection $d \in \mathcal{D}_{\text{global}}^{(k)}$, we define:

$$Q_d^t = s_d \cdot \hat{\text{IoU}}_d^t \quad \forall t \in [1, T_v] \tag{20}$$

where Q_d^t is the quality score of detection d in frame t, s_d is the confidence score of detection d, and T_v denotes the number of frames in video v.

We implement quality-based pseudo-label selection using a fixed quality threshold τ_{th}. For each detection $d \in \mathcal{D}_{\text{global}}^{(k)}$ across all videos:

$$\mathcal{S}_d^t = \begin{cases} 1 & Q_d^t \geq \tau^{(k)} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where $\mathcal{S}_d^t$ denote whether pseudo-label of detection d in frame t is selected. For each detection, if its results are not selected in any frames, we discard it:

$$\mathcal{D}_{\text{retained}}^{(k)} = \left\{ d_v \in \mathcal{D}_{\text{global}}^{(k)} \;\middle|\; \sum_{t=1}^{T_v} \mathcal{S}_d^t > 0 \right\} \tag{22}$$

where $\mathcal{D}_{\text{retained}}^{(k)}$ is the set of detections we retain in iteration k.

7. Additional qualitative visualizations

We provide additional qualitative results of our VIS model and quality predictor in Fig. 8 and Fig. 9.

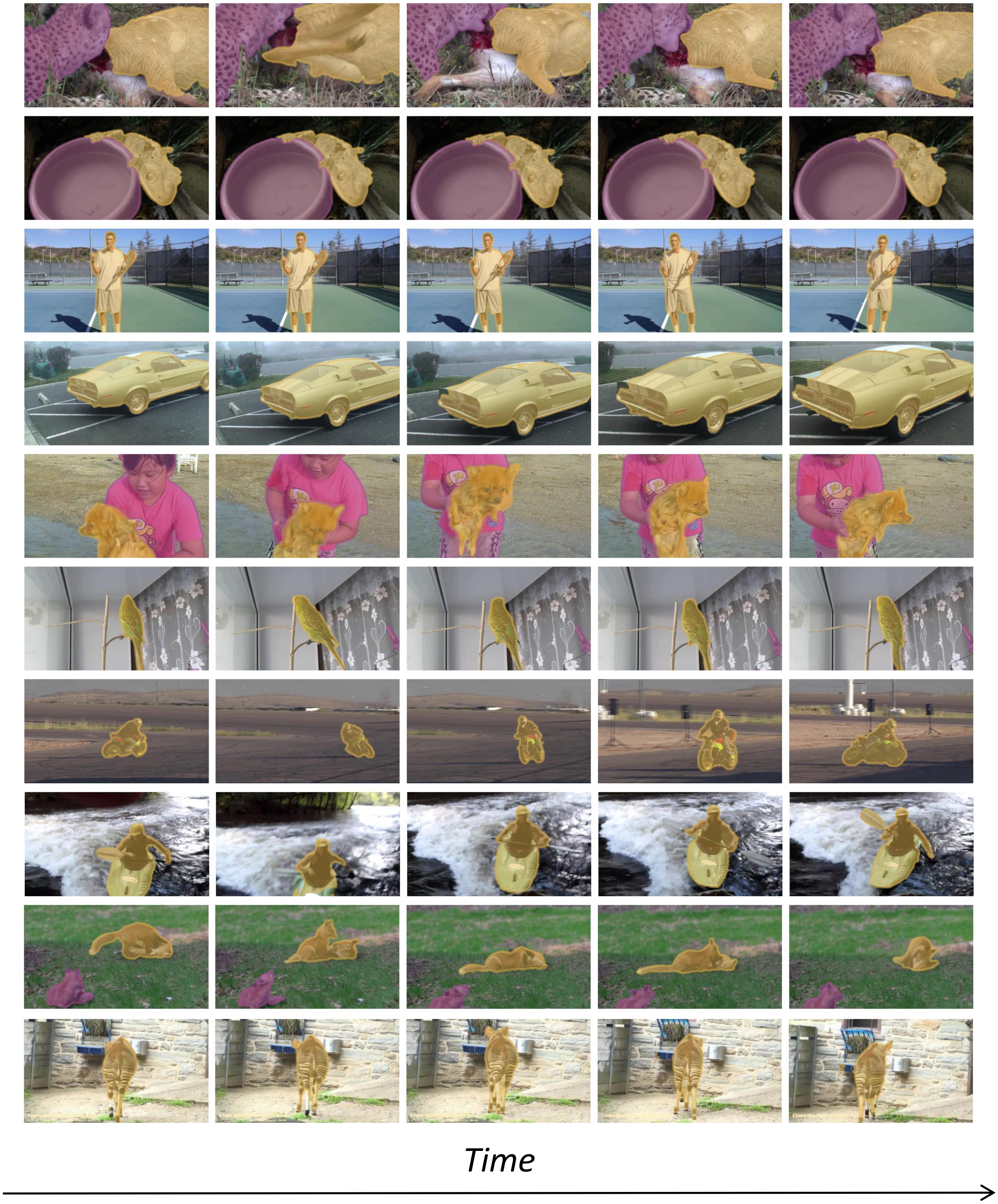

Figure 8. **Qualitative results of our VIS model on YouTubeVIS-2019 `val` split.**

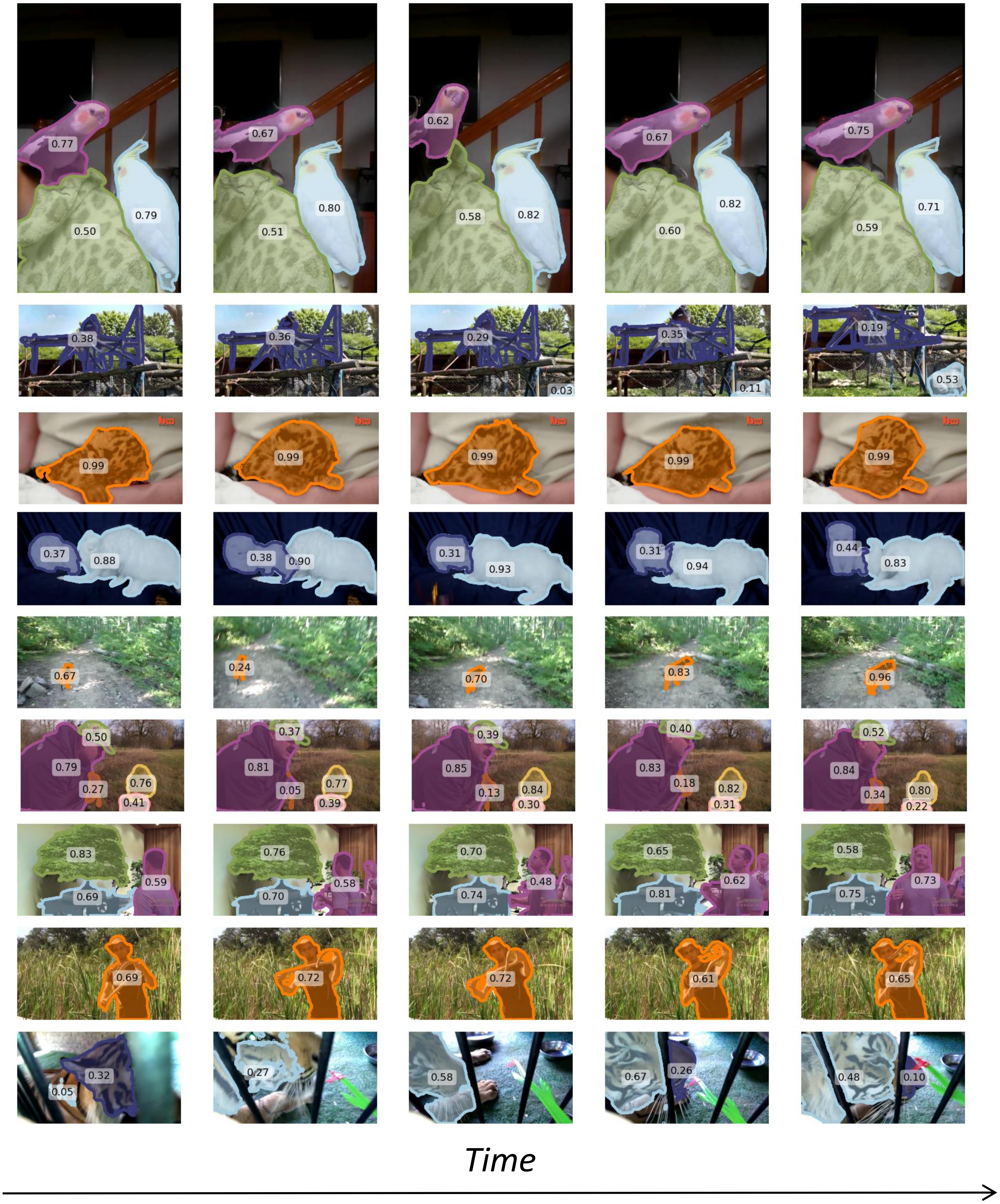

Figure 9. **Qualitative results of our quality predictor on YouTubeVIS-2019 `train` split.** The quality scores are shown in the center of each pseudo label.