Supplementary Material: Video Instance Segmentation Tracking with a Modified VAE Architecture

Chung-Ching Lin
IBM Research AI
cclin@us.ibm.com

Ying Hung
Rutgers University
yhung@stat.rutgers.edu

Rogerio Feris
IBM Research AI
rsferis@us.ibm.com

Linglin He
Rutgers University
lhe@stat.rutgers.edu

1. Mathematical Details

We describe more mathematical details of our variational inference with Gaussian Process latent variables.

Lemma 1 (Covariance under spatial correlation assumptions). *Assume that the latent variables $z = (z_1, \ldots, z_J)$ can be divided into k independent groups, within which the latent variables are correlated. Denote $(z_{m_1}, \ldots, z_{m_{d_m}})$ as the m^{th} group with d_m components, where $m = 1, \ldots, k$ and $\sum_{m=1}^{k} d_m = J$. Defining the correlation structure by $Corr(z_{m_i}, z_{m_j}) = \rho_m < 1$ when $i \neq j$ and $Corr(z_{m_i}, z_{n_j}) = 0$ when $m \neq n$, the determinant of the covariance matrix can be written as:*

$$|\Sigma| = \prod_{i=1}^{J} \sigma_i^2 \prod_{m=1}^{k} (1-\rho_m)^{d_m - 1}(\rho_m d_m + 1 - \rho_m). \quad (5)$$

Proof. The covariance matrix can be written as

$$\Sigma = \begin{bmatrix} D_1 & 0 & 0 & \ldots & 0 \\ 0 & D_2 & & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & D_k \end{bmatrix}, \quad (6)$$

where

$$D_m = \begin{bmatrix} \sigma_{m_1}^2 & \sigma_{m_1}\sigma_{m_2}\rho_m & \ldots & \sigma_{m_1}\sigma_{m_{d_m}}\rho_m \\ \sigma_{m_2}\sigma_{m_1}\rho_m & \sigma_{m_2}^2 & \ldots & \sigma_{m_2}\sigma_{m_{d_m}}\rho_m \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m_{d_m}}\sigma_{m_1}\rho_m & \sigma_{m_{d_m}}\sigma_{m_2}\rho_m & \ldots & \sigma_{m_{d_m}}^2 \end{bmatrix}. \quad (7)$$

With the spatial correlation assumption, it can be derived that

$$|D_m| = \prod_{i=1}^{d_m} \sigma_{m_i}^2 (1-\rho_m)^{d_m - 1}(\rho_m d_m + 1 - \rho_m). \quad (8)$$

Thus we have

$$\begin{aligned} |\Sigma| &= \prod_{m=1}^{k} \prod_{i=1}^{d_m} \sigma_{m_i}^2 (1-\rho_m)^{d_m - 1}(\rho_m d_m + 1 - \rho_m) \\ &= \prod_{i=1}^{J} \sigma_i^2 \prod_{m=1}^{k} (1-\rho_m)^{d_m - 1}(\rho_m d_m + 1 - \rho_m). \end{aligned} \quad (9)$$

□

Given the spatial correlation structure in Lemma 1, the corresponding D_{KL} divergence is derived in Theorem 2.

Theorem 2 (D_{KL} divergence under spatial correlation assumption). *Under the spatial correlation assumptions in Lemma 1 and the results in Equation (5), the D_{KL} divergence can be derived:*

$$\begin{aligned} &- D_{KL}(q_\phi(z|\xi,\varphi)||p_\theta(z|\varphi)) \\ =&\frac{1}{2}\sum_{m=1}^{k}(d_m - 1)\log(1-\rho_m) + \log(\rho_m d_m + 1 - \rho_m) \\ &+ \frac{1}{2}\sum_{j=1}^{J}(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2). \end{aligned} \quad (10)$$

Proof. It can be shown that

$$\begin{aligned} &\int q_\phi(z|\xi,\varphi)\log p_\theta(z|\varphi)dz \\ =&-\frac{J}{2}\log(2\pi) - \frac{1}{2}E_{q_\phi(z|\xi,\varphi)}z^T z \\ =&-\frac{J}{2}\log(2\pi) - \frac{1}{2}tr(\Sigma + uu^T), \end{aligned} \quad (11)$$

and

$$\begin{aligned} &\int q_\phi(z|\xi,\varphi)\log q_\phi(z|\xi,\varphi)dz \\ =&-\frac{J}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| \\ &-\frac{1}{2}\int (z-u)^T\Sigma^{-1}(z-u)q_\phi(z|\xi,\varphi)dz \\ =&-\frac{J}{2}\log(2\pi) - -\frac{1}{2}\log|\Sigma| - \frac{J}{2}. \end{aligned} \quad (12)$$

Therefore, plugging in Equations (11) and (12), the D_{KL}

divergence can be derived as

$$
\begin{aligned}
& - D_{KL}(q_\phi(z|\xi,\varphi)||p_\theta(z|\varphi)) \\
& = \int q_\phi(z|\xi,\varphi)(\log p_\theta(z|\varphi) - \log q_\phi(z|\xi,\varphi))dz \\
& = \frac{1}{2}(J + \log|\Sigma| - tr(\Sigma + uu^T)) \\
& = \frac{1}{2}\sum_{j=1}^{J}(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2) \\
& + \frac{1}{2}\sum_{m=1}^{k}(d_m - 1)\log(1-\rho_m) + \log(\rho_m d_m + 1 - \rho_m).
\end{aligned} \tag{13}
$$

□

2. Evaluation Metrics

We adopt the evaluation metrics proposed in Voigtlaender *et al*. [1]. The metrics that are used to measures the quality of the segmentation as well as the consistency of the predictions over time are listed in Table 3. Different from bounding box detection, where a ground truth box may overlap with several predicted boxes, in instance segmentation, since each pixel is assigned to at most one instance, only one predicted mask can have an Intersection over Union (IoU) larger than a given threshold with a given ground truth mask.

We summarize the evaluation metrics defined in [1]. Given a set of estimated mask $H = \{h_1, \ldots, h_K\}$ and a set of ground truth masks $M = \{m_1, \ldots, m_N\}$, a mapping $c(h)$ from an estimated mask h to ground truth mask $m \in M$ can be defined using mask-based IoU as:

$$
c(h) = \begin{cases} \arg\max_{m \in M} IoU(h,m), & \text{if } \max_{m \in M} IoU(h,m) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{14}
$$

A soft version $\widetilde{TP}$ of the number of true positives is defined in [1]:

$$
\widetilde{TP} = \sum_{h \in TP} IoU(h, c(h)). \tag{15}
$$

The multi-object tracking and segmentation accuracy (MOTSA) is defined as a mask IoU based version of the box-based MOTA metric, i.e.

$$
MOTSA = 1 - \frac{|FN| + |FP| + |IDS|}{M} \tag{16}
$$

$$
= \frac{|TP| + |FP| + |IDS|}{M} \tag{17}
$$

and the mask-based multi-object tracking and segmentation precision (MOTSP) as

$$
MOTSP = \frac{\widetilde{TP}}{|TP|} \tag{18}
$$

The soft multi-object tracking and segmentation accuracy (sMOTSA) proposed in [1] is defined as:

$$
sMOTSA = \frac{\widetilde{TP} + |FP| + |IDS|}{M}, \tag{19}
$$

which accumulates the soft number $\widetilde{TP}$ of true positives instead of counting how many masks reach on IoU of more than 0.5. sMOTSA therefore measures segmentation as well as detection and tracking quality.

3. Qualitative Evaluation

In this section, we present additional qualitative results:

1. Figure 4 shows an example of failure case. Our model is able to use spatial interdependency and motion continuity to reduce false negatives and improve VIST performance in most challenging cases. However, it still has limitations in some extremely challenging cases, e.g., at Frame 121, most of the pedestrian's body parts are occluded by the traffic light pole.
2. Figure 5, 6, and 7 show more qualitative results of our proposed method on the KITTI MOTS [1], MOTSChallenge [1], and YouTube-VIS [2] datasets.

References

[1] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 2

[2] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

Table 3: Evaluation metrics for video instance segmentation tracking.

Name	Definition
sMOTSA↑	The soft Multiple Object Tracking and Segmentation Accuracy takes into account false positives, missed targets and identity switches
MOTSA↑	The Multiple Object Tracking and Segmentation Accuracy takes into account false positives, missed targets and identity switches
MOTSP↑	The mask-based Multiple Object Tracking and Segmentation Precision is simply the average IoU between true and estimated targets
True positives (TP) ↑	Number of correctly matched masks.
Soft true positives ($\widetilde{TP}$) ↑	Sum of IoU of correctly matched masks.
False positives (FP) ↓	Number of predicted masks not assigned to any ground truth mask.
False negatives (FN) ↓	Number of ground truth masks not matched by any estimated masks.

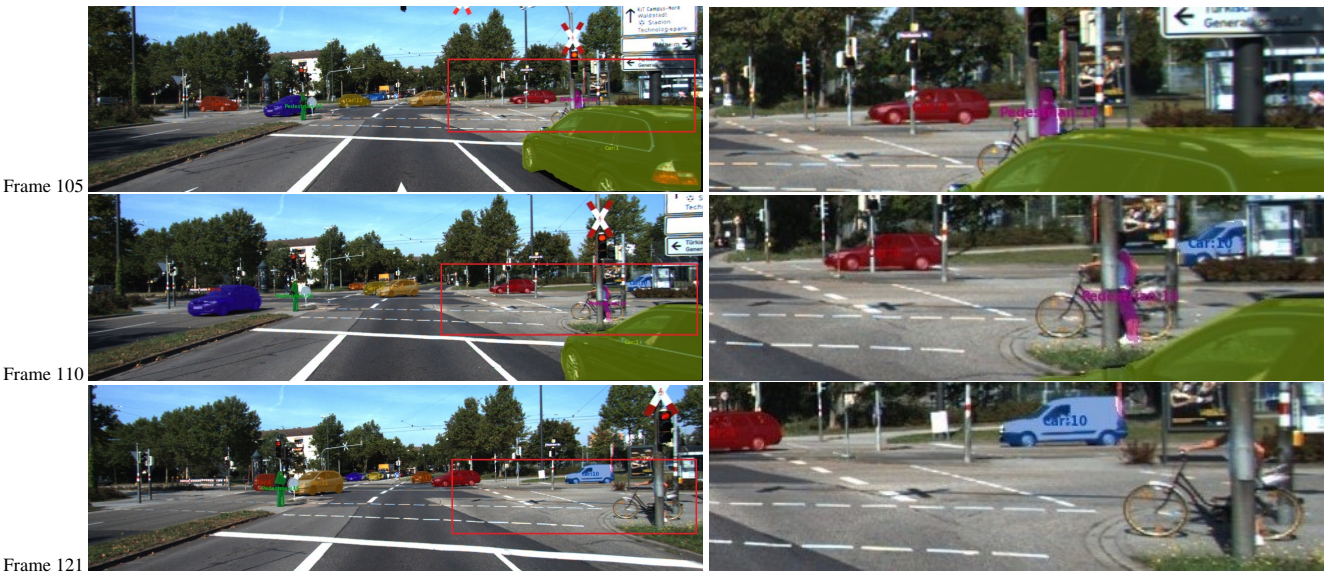

Figure 4: An example of failure case. Left column shows the original frames of video sequence 02 in KITTI MOTS dataset. The regions highlighted with red rectangles are further zoomed and shown in the right column. Our model is able to use spatial interdependency and motion continuity to reduce false negatives and improve VIST performance in most challenging cases. However, it still has limitations in some extremely challenging cases, e.g., in the above Frame 121, most of the pedestrian's body parts are occluded by the traffic light pole.

Figure 5: The visualization of our proposed method on KITTI MOTS dataset.

Figure 6: The visualization of our proposed method on MOTSChallenge dataset.

Figure 7: The visualization of our proposed method on YouTube-VIS dataset.