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

Figure 1: Our model classifies, localizes, segments, and tracks all instances of predefined object classes with consistent assigned identities.

Abstract

We propose a modified variational autoencoder (VAE) architecture built on top of Mask R-CNN for instance-level video segmentation and tracking. The method builds a shared encoder and three parallel decoders, yielding three disjoint branches for predictions of future frames, object detection boxes, and instance segmentation masks. To effectively solve multiple learning tasks, we introduce a Gaussian Process model to enhance the statistical representation of VAE by relaxing the prior strong independent and identically distributed (iid) assumption of conventional VAEs and allowing potential correlations among extracted latent variables. The network learns embedded spatial interdependence and motion continuity in video data and creates a representation that is effective to produce high-quality segmentation masks and track multiple instances in diverse and unstructured videos. Evaluation on a variety of recently introduced datasets shows that our model outperforms previous methods and achieves the new best in class performance.

1. Introduction

In recent years, there has been great progress in the area of automatic video understanding. Classic video tasks are centered around understanding what objects are doing and their actions. This paper considers an emerging video understanding task: Video Instance Segmentation Tracking (VIST), which aims to classify, localize, segment, and track all instances of object classes throughout a video and yield pixel-wise object labels [59, 66]. This task provides a more natural understanding of the video scenes and is more desirable for applications that require detailed pixel-level information, such as autonomous driving and video editing.

The VIST task is different from traditional video object tracking and video object segmentation. Video object tracking [1, 4, 35, 54, 63, 65, 68] uses bounding boxes to identify the target objects and to estimate their positions in the subsequent frames. Yet, in many scenarios with heavy occlusions, simple rectangular bounding boxes fail to properly represent objects. The VIST task produces binary segmentation masks and pixel-level tracking results. Furthermore, if objects are occluded, or out of the scene for a couple of frames before reappearing, the instance identities are maintained. Traditional video object segmentation can coarsely

be separated into two groups: semi-supervised and unsupervised. VIST differs from both of them. In semi-supervised mode [9, 60, 46, 12, 21], the initial masks for objects of interest are provided in the first frame. In unsupervised mode [22, 70, 40, 39], only salient objects are to be tracked. In the proposed VIST task, initial masks are not available and objects to be tracked are set by predefined classes.

Despite the remarkable progress achieved with CNNs, VIST is still challenging when applied to real world environments. To address the VIST task, Voigtlaender *et al*. [59] propose TrackR-CNN, which extends Mask R-CNN [18] with 3D convolutions to incorporate temporal information and adds an association head to link object identities over time. Similarly, Yang *et al*. [66] propose MaskTrack R-CNN, which introduces a new tracking branch to Mask R-CNN to jointly perform the detection, segmentation and tracking tasks. TrackR-CNN and MaskTrack R-CNN are both nicely designed models and demonstrate promising directions of adapting Mask R-CNN with an association head for tracking. However, both methods assume that Mask R-CNN is effective in producing well-localized bounding boxes and accurate segmentation results. In highly diverse and unstructured videos, visual objects are often subject to partial or even full occlusion, deformation, pose variation, and, in many cases, objects have similar appearance and are hard to be isolated from a cluttered background. Thus, there is a great possibility that object detections are ill-initialized, which, in turn, degrade the precision of object masks predicted within bounding boxes and tracking results produced by linking masks. In other words, the mechanism of directly linking Mask R-CNN segmentation masks across frames via an association head faces an inherent limitation: the model has difficulties in handling false negative proposals, leading to inferior performance.

In this paper, we propose a variational autoencoder (VAE) modification that builds on top of Mask R-CNN, to tackle the VIST problem. We note that the spatial interdependence and motion continuity across frames provide a supportive context that allows a video model to better infer what is happening next. We adapt a VAE architecture to capture spatial and motion information shared by all instances, and generate attentive cues to reduce false negative mask predictions. By forcing the network to solve multiple learning tasks, we induce a representation within the network which guides well the video instance segmentation tracking task and produce high quality segmentation masks. Figure 1 illustrates sample experimental results of our method on MOTS [59] and YouTube-VIS [66] datasets.

Our contributions are summarized as follows. (1) Our multi-task network architecture and training scheme have been carefully designed to take advantage of both spatial and motion cues. It achieves the new best in class performance with the same network on the recently released KITTI MOTS, MOTS Challenge [59] and YouTube-VIS [66] datasets. (2) We introduce a Gaussian Process model to enhance the statistical representation of VAE by relaxing the prior strong independent and identically distributed *(iid)* assumption of conventional VAEs and allowing correlations among extracted latent variables. The spatial interdependence is encoded by the modified VAE, which plays a crucial role in generating valid and errorless instance segmentation.

2. Related Work

Image Instance Segmentation. The instance segmentation task [43, 18, 45, 10] is closely related to object detection and semantic segmentation. A mainstream framework to solve this task is to augment a detector network with a branch to predict object masks within bounding boxes or region proposals. He et al. propose Mask R-CNN [18] that extends the Faster R-CNN framework with a mask head and achieves state-of-the-art performance. The idea is further developed by PANet [45] and Mask Scoring R-CNN [24] which outperform competing methods on COCO dataset. These methods have achieved impressive performance on localizing objects of interest at pixel level in images, but they are not directly applicable to the VIST task. In the VIST task, object instances not only are segmented and represented by masks in each video frame, but also are tracked with the same corresponding identities throughout a video.

Visual Object Tracking. The typical visual object tracking (VOT) [1, 4, 35, 54, 63, 65, 68] is based on bounding boxes and usually does not provide accurate object contours. VOT [30], MOT Challenge [34], PETS [14] and KITTI [16] are popular datasets used to evaluate VOT performance. Existing methods for this task roughly fall into two categories, namely detection-free and detection-based. Detection-free tracking methods [5, 58, 37, 62, 36, 69] track objects given a manual initialization of a fixed number of objects in the first frame. They cannot deal with new objects appearing in the middle of a video sequence. Detection-based tracking [27, 67, 63, 20, 49, 13, 4, 61, 32, 42], on the other hand, generally requires objects being detected followed by a tracker that links the detection regions to form trajectories of the targets. In order to resolve ambiguities in linking object detections, many efforts have explored this problem with data association approaches [27, 67], such as Markov decision process [63], event aggregation [20], greedy algorithm [49], and attentional correlation filters [13].

Video Object Segmentation. Video object segmentation (VOS) aims at segmenting and tracking objects in videos, but does not require recognition of object categories. Popular datasets for the VOS task include DAVIS [50], SegTrackV2 [38] and YouTube-VOS [64]. In general, VOS has two major categories: semi-supervised and unsupervised.

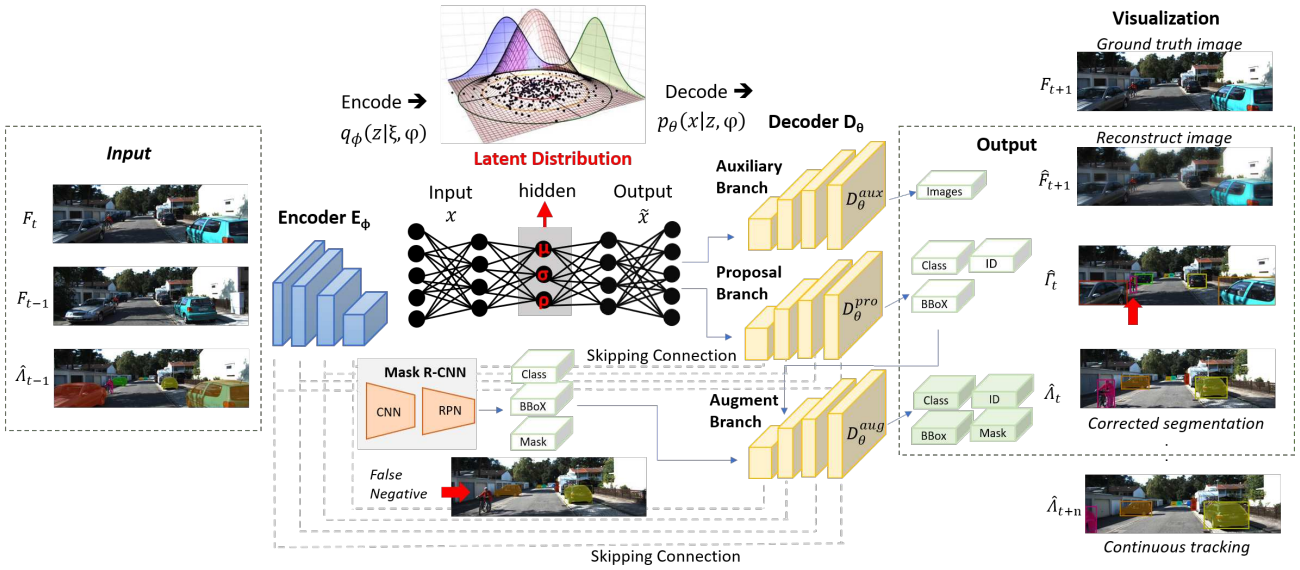


Figure 2: Illustration of proposed framework for video instance segmentation tracking.

In semi-supervised scenario, ground-truth masks are given in the first mask and tracked through the rest of frames of the sequence. Typically, spatial-temporal graph and CNN based methods are investigated. Spatial-temporal graph methods rely on two important cues: (1) object representation of graph structure, e.g., pixels [57, 47], superpixels [51, 25], object patches [3] and (2) spatial-temporal connections, e.g., spatial-temporal lattices [47], nearest neighbor fields [2, 15], and mixture of trees [8]. Some of the CNN-based methods [21, 41] employ Recurrent Neural Networks with optical flow to capture the temporal coherence of object motion and propagate information between frames. Another line of methods [11, 23, 53] formulates VOS as a pixel-wise matching problem. Other approaches [9, 60] learn an appearance model to perform pixel-level detection and segmentation of objects at each frame.

In the unsupervised scenario [22, 70, 40, 39, 55, 33], the task is to segment salient foreground moving objects in a fully automatic way. Motion patterns, (e.g. optical flow [56] and long-term trajectory [7]) are usually used as main sources of information. Due to the lack of guidance from object masks, most of the unsupervised methods cannot segment a specific object if there exists motion ambiguity between different instances and dynamic background.

In summary, VIST has some common challenges as VOT and VOS, but differs in several aspects, for example, no first mask is given as guidance, all object instances belonging to a set of predefined classes are needed to be classified, segmented and tracked with the consistent assigned identities throughout a video, and the final outputs are precise pixel-level masks without any overlapping pixels between masks. To evaluate VIST performance, the masks of all instances of a predefined category set and the instance identities across frames should be labeled. Thanks to the KITTI MOTS, MOTS Challenge, and YouTube-VIS datasets recently introduced by [59] [66], the effectiveness of the proposed method could be evaluated.

3. Method

3.1. Overview

We adopt a VAE architecture which consists of a probabilistic variable to describe an observation in latent space. Our method is illustrated in Figure 2. Specifically, our VAE architecture includes one encoder and three decoders, which yields three parallel branches, namely auxiliary branch, proposal branch and augment branch. The skip connection scheme is applied between the down-sampling encoder and up-sampling decoders in the proposal branch and augment branch for information preservation. These three branches share the same hidden feature layers but perform different tasks. The *auxiliary branch* takes frame-level video inputs and learns to predict future frames. The goal of this branch is to guide the network to learn finer representations and increase the amount of meaningful semantic information encoded in the latent space. The *proposal branch* summarizes and outputs object-level information for connecting objects over time. It also provides attentive cues to reduce false negatives in the augment branch. The *augment branch* aggregates pixel-level features extracted from different layers in the VAE encoder and the Mask R-CNN network. The low-level features are rich in spatial details, and the high-level features contain more semantic information. By combining these extracted features with outputs from the proposal

branch and the Mask R-CNN network, this branch produces final instance classifications, identities, detection boxes and segmentation masks.

3.2. Unified Variational Autoencoder

We consider a video sequence F which consists of T frames F_t, $t \in \{1, ..., T\}$, with N instances belonging to a predefined category label set C. Our variational inference network follows an encoding-decoding scheme, consisting of four components: an encoder E_ϕ and three conditional decoders D_θ^{aux} (Auxiliary branch), D_θ^{pro} (Proposal branch), D_θ^{aug} (Augment branch). The variational network takes the current observation $\xi_t = [F_t, F_{t-1}, \Lambda_{t-1}]$ as an input to perform multi-task learning: predict the future frame $\hat{F}_{t+1}$ in D_θ^{aux}, generate a set of detection box predictions $\hat{\Gamma}_t$ in D_θ^{pro}, and estimate a set of instance segmentation masks $\hat{\Lambda}_t$ in D_θ^{aug}. We denote $\hat{\Gamma}_t = \{b_{i,t}\}_{i=1}^{n_b}$ and $\hat{\Lambda}_t = \{m_{i,t}\}_{i=1}^{n_m}$, where $b_{i,t}$ and $m_{i,t}$ are the detection box prediction and the segmentation mask for instance i at frame t respectively. n_b and n_m are the number of detected object instances at frame t in D_θ^{pro} and D_θ^{aug}. More specifically, the encoder E_ϕ first maps the current observation ξ_t to a latent variable z and a spatial prior φ. The conditional decoders D_θ computes z and φ to estimate the output $\hat{\chi}_t = [\hat{F}_{t+1}, \hat{\Gamma}_t, \hat{\Lambda}_t]$.

We formulate our encoder to describe a probability distribution for each latent attribute. In conventional VAE models, latent encoding variables are assumed to be identically and independently distributed (*iid*) across both latent dimensions and samples, which is not realistic in many problems with high dimensional inter and intra-data correlation. For example, in a video sequence, it is reasonable to expect that the frames that were taken adjacently would exhibit similar latent representations. With this rationale, we propose to relax the strong prior *iid* assumption of standard VAE, and allow correlation among latent variables to model the spatial interdependence observed in video data. This encourages the encoder to capture better representations of the underlying data distribution across videos frames.

3.2.1 Conditional Variational Bound

Given that z is a latent variable and φ is a conditional prior, one way of learning the decoder $p_\theta(\chi_t|z, \varphi)$ is to use a variational autoencoder [29]. But, instead of using the same data for input and target output, we take the current observations ξ as input and targeted estimation χ as output. Additionally, we add another constraint on the network, using a conditional prior φ extracted from ξ to preserve spatial information.

The decoder network D_θ estimates the parameters of the distribution $p_\theta(\chi_t|z, \varphi)$. To learn the decoder, we need to maximize the log-likelihood of observed data ξ and marginalize out the latent variables z and φ. To avoid the intractable integral, an approximate posterior $q_\phi(z|\xi, \varphi)$ is introduced to obtain the ELBO from Jensen's inequality [29],

$$\log p_\theta(\chi_t|\xi) \geq \mathbb{E}_q \log \frac{p_\theta(\chi_t|z, \varphi)p_\phi(z|\varphi)}{q_\phi(z|\xi, \varphi)}. \quad (1)$$

The loss function for training these models follows directly from Equation 1 and has the form: $\mathcal{L}(\chi_t, \theta, \phi) = -D_{KL}(q_\phi(z|\xi, \varphi)||p_\theta(z|\varphi)) + \mathbb{E}_{q_\phi(z|\xi,\varphi)}[\log p_\theta(\chi_t|z, \varphi)]$.

The loss function is made up of two parts: a D_{KL} divergence and a log likelihood part. D_{KL} divergence part is latent loss, which can be understood as a distance between the distribution $q_\phi(z|\xi, \varphi)$ and a prior distribution for z. By minimizing this distance, we are really avoiding that $q_\phi(z|\xi, \varphi)$ departs too much from its prior, acting as a regularization term. The second part is decoding loss, which measures how accurately the network constructed the semantic output χ_t by using the distribution $p_\theta(\chi_t|z, \varphi)$, that is, it is a distance between $\hat{\chi}_t$ and χ_t.

3.2.2 Variational Inference with Gaussian Process Latent Variables

Video data has very strong spatial correlation within and among frames; however, the conventional VAEs impose a strong assumption that the latent variables are all independent and identically distributed. To relax this assumption, we propose a new scheme assuming the prior $p_\theta(z|\varphi)$ following $N(0, \mathbf{I})$ and the latent variables to be realization from a constant mean Gaussian process denoted by $GP(u, \Sigma)$, where u denotes the variational mean and Σ is the covariance function accommodating the potential spatial correlation. In Lemma 1, the spatial correlation structure is defined and the determinant of the corresponding covariance matrix is derived.

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:*

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

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 (2), the D_{KL} divergence can be derived:*

$$
\begin{aligned}
&- D_{KL}(q_\phi(z|\xi,\varphi)||p_\theta(z|\varphi)) \\
=&\frac{1}{2}\sum_{m=1}^{k}(d_m-1)\log(1-\rho_m)+\log(\rho_m d_m+1-\rho_m) \\
&+\frac{1}{2}\sum_{j=1}^{J}(1+\log(\sigma_j^2)-\mu_j^2-\sigma_j^2). \qquad (3)
\end{aligned}
$$

To optimize the KL divergence derived in Theorem 2, we apply a reparameterization trick [29]: instead of the encoder generating a vector of real values, it generates a vector of means, a vector of standard deviations and a vector of correlations. When decoding from latent state, we sample from the Gaussian Process with their mean and covariance matrix, and use that as our latent vector z. This constraint forces the encoder to be very efficient, creating information-rich latent variables, and improves the generalization of our network to tolerate noise from various type of video contents.

Here, we extract and preserve spatial information φ from current observation ξ to enhance estimation of instance segmentation. In practical cases, a VAE network tends to produce more blurry images. We thus employ skip connections where the pooling operations in the contracting path (encoding) are mirrored by upsampling operations in the symmetry expanding (decoding) path. The skip connections from earlier layers in the network could provide the necessary spatial details in order to reconstruct accurate shapes for instance segmentation. Also, the symmetric expanding path enables precise localization. In our case, φ is the features from different layers in the encoder network and is fed into the corresponded layers with the same scale in the decoder network.

3.3. Decoder Branches

3.3.1 Auxiliary Branch

In VAEs, there might exist bypassing connection between the encoder and the decoder if the network is not designed properly. Specifically, if the decoder has a direct and deterministic access to the source, the latent variables z might not capture much information so that the VAE does not play an effective role in the process. To avoid this, we include an auxiliary branch, which plays as a supportive role to guide the model towards increasing the amount of semantic information encoded in the latent space and creating information-rich latent variables.

The training objective of this branch is to reconstruct the future frame F_{t+1} given the current observation ξ_t. The reconstruction loss is measured by mean square error (MSE) between the predicted and ground truth images. The auxiliary branch is implemented with n residual blocks, following the architecture proposed in [19] without batch normalization. We use strided convolution with stride of 2 after each residual block to down-sample the inputs until a bottleneck layer, and we utilize subpixel convolution [52] to perform the up-sampling between two consecutive residual blocks. All convolutional layers consist of 3x3 filters. The following two branches have similar architecture.

3.3.2 Proposal Branch

The goal of this branch is to summarize and output object-level information, which provides attentive cues to reduce false negatives in the augment branch. Let $\mathcal{I}$ denote the set of instances detected and identified at frame F_{t-1}. Each instance $i \in \mathcal{I}$ consists a detection box $b_{i,t-1}$, a segmentation map $m_{i,t-1}$, a classification $c_{i,t-1}$ and an identity $id_{i,t-1}$. For each object instance i, its localization at frame t are estimated by the Decoder D_θ^{pro} in this branch. Following the spirit of Mask R-CNN [21], we use the detection box $b_{i,t-1}$ as a region of interest (RoI), and apply RoIAlign to extract multi-scale feature maps and locate the relevant areas. The extracted features are passed through two fully connected layers for bounding box regression. We thus obtain the detection box prediction $\hat{b}_{i,t}^{pro}$. The detection box keeps the same identity as the corresponding instance. The bounding box loss is measured by a weighted smooth L1 loss [17] for backward propagation.

3.3.3 Augment Branch

We aim to produce high-quality instance segmentation masks and reliable object tracks in this branch. We first concatenate features from difference sources, match detection outputs from the proposal branch and the Mask R-CNN network, and then produce the final classification, identity, detection box and segmentation mask for each instance.

Considering the features from multiple sources are strong complements to existing box and mask features, we incorporate the features at different levels of our VAE encoder and Mask R-CNN network for better feature presentations. With our network design, the combined features contain more spatial details and motion information, and thus are more discriminative on cluttered background. The new feature map x^{Aug} can be generated as

$$
x_i^{Aug} = x_i^E \oplus x_i^{Mask}, \qquad (4)
$$

where $\oplus$ denotes concatenation, x_i^E and x_i^{Mask} are the i-th scale feature maps from Encoder E_ϕ and Mask R-CNN backbone respectively. The concatenated multiple-scale features are fed into the Decoder D_θ^{aug}.

The Decoder D_θ^{aug} takes instance detection box outputs from the Proposal branch and the Mask R-CNN network. Let $\hat{b}_{i,t}^{pro}$ denotes the detection box prediction for instance i at frame t from proposal branch, and $\hat{b}_{j,t}^{mask}$ denotes the

detection box output for instance j at frame t from Mask R-CNN network. We first match $\hat{b}_{i,t}^{pro}$ and $\hat{b}_{j,t}^{mask}$. The matching cost between bounding boxes is defined as:

$$c_{ij} = 1 - \Omega(\hat{b}_{i,t}^{pro},\ \hat{b}_{j,t}^{mask}), \tag{5}$$

where $\Omega(\cdot,\cdot)$ is the intersection over union (IoU) ratio of bounding boxes. We match the bounding boxes only if their IoU ratio is greater than ϵ. After computing all matching costs between $\hat{b}_{i,t}^{pro}$ and $\hat{b}_{j,t}^{mask}$, we find the optimal set of matching pairs using the Hungarian algorithm [31]. Then we create a new rectangle detection box by the union of the matched box pair. The new bounding box takes the same identity as corresponding $\hat{b}_{i,t}^{pro}$. The unmatched bounding box prediction $\hat{b}_{i,t}^{pro}$ keeps its original identify. The unmatched detection box $\hat{b}_{j,t}^{mask}$ is considered a newly-appearing object instance and is assigned a new identity.

For each detection box with positive RoI, we apply RoIAlign to extract feature map, perform object class and bounding box regression, and use a fully convolutional network (FCN) to generate a pixel-level mask. The new instance segmentation mask is denoted by $m_{i,t}$ and is linked to the ith instance track to form the track $\mathcal{T}_{i,t} = \{..., m_{i,t-1}, m_{i,t}\}$. We use average binary cross-entropy to measure mask loss [18].

4. Experiments and Analysis

4.1. Experimental Setup.

Dataset. We evaluate the proposed method on the newly introduced KITTI MOTS, MOTSChallenge [59] and YouTube-VIS [66] datasets for video instance segmentation tracking. The objects in these datasets have consistent instance identity labels across frames. KITTI MOTS dataset focuses on videos from vehicle-mounted cameras. It contains 8,008 frames in 21 scenes, with 26,899 annotated cars and 11,420 annotated pedestrians. MOTS Challenge dataset presents pedestrians in crowded scenes. It contains 2,862 frames with 26,894 annotated pedestrians. YouTube-VIS dataset [66] contains Internet videos, covering 40 categories, such as animals, cars accessories and human. However, the annotations in the released testing and validation sets do not include object instance identities, making it unsuitable to train and test VIST methods. Thus, we randomly split the YouTube-VIS training set into 2038 training videos and 200 test videos. There are some differences in these datasets. KITTI MOTS and MOTSChallenge dataset contain less classes, but have much longer videos with more objects that frequently disappear and reappear in the scenes.

Evaluation Metrics. Following TrackR-CNN [59], we report evaluation metrics: soft multi-object tracking and segmentation accuracy (sMOTSA), multi-object tracking and segmentation accuracy (MOTSA), and multi-object tracking and segmentation precision (MOTSP). We also report true positive (TP), false positive (FP) and false negative (FN). Among these metrics, sMOTSA is the recommended primary metric to measure performance, as it considers segmentation as well as detection and tracking quality [59].

Implementation Details. We use a ResNet-101 [19] backbone for Mask R-CNN, and pre-train it on COCO [44] and Mapillary [48] for experiments on KITTI MOTS and MOTSChallenge datasets, and pre-train it on COCO for experiments on YouTube-VIS dataset. We implement our model in PyTorch [26] and train it with 4 GeForce RTX 2080 Ti GPUs. Each batch has 8 images (each GPU holds 2 images). We train the model on the target datasets for 20 epochs with a learning rate of 1.5×10^{-4} using Adam optimizer [28].

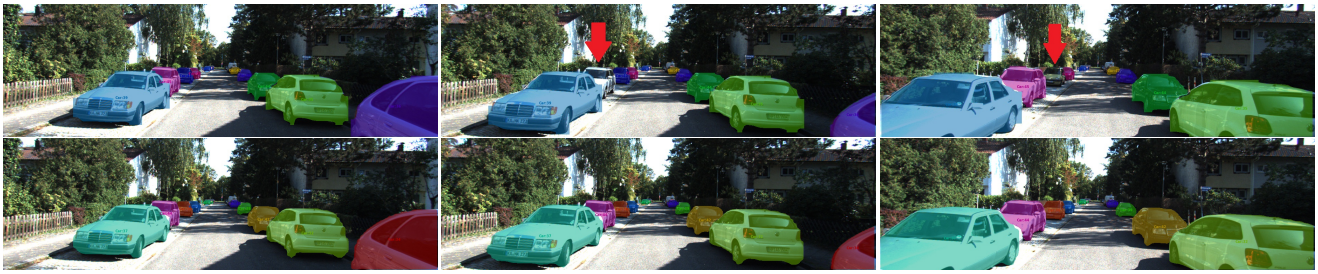
We generate image-level combined segmentation mask Λ_{t-1}^{l} using weighted instance masks in different categories. Λ_{t-1}^{l} is concatenated with images F_t^l and F_{t-1}^l as an input to the encoder. The long edge and short edge of images are resized to 1333 and 800 pixels respectively without changing the aspect ratio.

4.2. Quantitative Results

We perform a thorough comparison of the proposed method to the state of the arts, MaskTrack R-CNN [66] and TrackR-CNN [59], and report the video instance segmentation and tracking performance on Table 1. Mask R-CNN+IT denotes a basic baseline of linking MaskR-CNN output masks by IoU matching with IoUTracker [6]. For a fair comparison, we re-train MaskTrack R-CNN [66] and TrackR-CNN [59] on our hardware platform, using the same ResNet-101 backbone and following their training and evaluation protocol. An important difference between MaskTrack R-CNN [66] and TrackR-CNN [59] is that MaskTrack R-CNN allows segmentation masks to overlap, while TrackR-CNN requires no overlapping masks. In reality, each pixel should only belong to one object. Thus, in VIST task, we define that there must not have any overlapping pixels between masks; each pixel is only assigned to one object instance.

KITTI MOTS. Table 1 (a)(b) show the experimental results on KITTI MOTS dataset [59]. In this dataset, there are two categories: cars and pedestrians. The results show that our method achieves promising improvement against the state-of-the-art methods for most metrics, demonstrating the superiority of our method. Overall, we achieve better performance in terms of sMOTSA, MOTSA, MOTSP, TP, FP and FN. Specifically, our method noticeably reduces the number of false negative (FN) and increases sMOTSA, which is used to evaluate overall detection, segmentation and tracking quality.

MOTSChallenge. We report the results on MOTS Challenge in Table 1 (c). MOTS Challenge includes videos with



(a) Sample results on the KITTI MOTS dataset. This example shows multiple cars partially occluded by other cars.

(b) Sample results on the MOTS Challenge dataset. This example shows a group of pedestrians in the crowded scene.

Figure 3: Qualitative comparisons between TrackR-CNN[59] and the proposed method on challenging cases. In both cases, each row shows the same output frames for TrackR-CNN[59] (top row) and ours (bottom row). Red arrows indicate false negatives. Both cases show the proposed method is able to reduce false negatives, produce instance masks and correctly maintain their identities in the cluttered scenes. Best viewed on screen.

Table 1: Quantitative comparisons on KITTI MOTS, MOTSChallenge and YouTube-VIS dataset.

Method	sMOTSA	MOTSA	MOTSP	TP ↑	FP ↓	FN ↓
(a) KITTI MOTS Dataset [59] - Cars						
Mask R-CNN [18]+IT[6]	74.9	85.8	85.1	7,109	148	920
MaskTrack R-CNN [66]	75.5	86.1	86.5	7,135	140	894
TrackR-CNN [59]	76.2	87.8	87.2	7,276	134	753
Ours	**77.6**	**88.8**	**87.7**	**7,355**	**130**	**674**
(b) KITTI MOTS Dataset [59] - Pedestrians						
Mask R-CNN [18]+IT[6]	44.6	63.8	74.1	2,479	295	868
MaskTrack R-CNN [66]	45.9	64.6	74.9	2,497	280	850
TrackR-CNN [59]	46.8	65.1	75.7	2525	267	822
Ours	**49.7**	**67.6**	**77.0**	**2,607**	**251**	**740**
(c) MOTSChallenge Dataset [59]						
Mask R-CNN [18]+IT[6]	48.6	65.5	77.6	19,676	1,939	7,218
MaskTrack R-CNN [66]	50.5	66.7	78.3	19,882	1,882	7,012
TrackR-CNN [59]	52.1	67.5	79.5	20,255	1,702	6,639
Ours	**59.5**	**71.5**	**84.7**	**21,253**	**1,537**	**5,641**
(d) YouTube-VIS Dataset [66]						
Mask R-CNN [18]+IT[6]	33.7	46.4	78.8	2,751	790	596
MaskTrack R-CNN [66]	34.1	47.2	78.7	2,767	789	580
TrackR-CNN [59]	34.6	48.3	79.8	2,801	**778**	546
Ours	**35.1**	**50.4**	**80.8**	**2,866**	785	**481**

pedestrians in highly occluded scenes. In general, pedestrians are one of the most challenging categories for instance segmentation and tracking. Table 1 (c) shows our method outperforms other methods with noticeable margin. For example, our method achieves sMOTSA of 59.5%. TrackR-CNN achieves the second-best sMOTSA of 52.1%. This implies the proposed method can overcome the difficulty and handle highly challenging videos better. We believe the significant performance difference lies in our modified variational autoencoder architecture is designed to capture the information of spatial interdependency and motion continuity in video data, and to compensate the insufficiency of adapting Mask R-CNN with an association method for VIST task.

YouTube-VIS. The results on YouTube-VIS is reported in Table 1(d). This is a challenging dataset which contains 40 categories, with many similar categories, such as ape and monkey. Also, most of the videos in this dataset are short videos (e.g. 100 frames) and are only labeled with interval of 5 frames. Thus, all methods yield more false positives and lower sMOTSA. Our method still shows its strength in most metrics. This indicates our proposed method could provide valuable complementary information for the localization of object instances and correct ill-initialized mask generation.

Qualitative Evaluation. Overall, TrackR-CNN demonstrates the second-best performance in the quantitative evaluation. We further present the qualitative outputs of TrackR-CNN and our proposed method in Figure 3. Each

Table 2: Ablation study of proposed method on KITTI MOTS dataset.

Method	Component-wise[1]			Branch-wise[2]			Cars			Pedestrians		
	Skipping Connection	VAE ρ Correlation	Variational Inference	Auxiliary Branch	Proposal Branch	Augment Branch	sMOTSA	MOTSA	MOTSP	sMOTSA	MOTSA	MOTSP
(a) Disable Skipping Connections		✓	✓	-	-	-	75.4	87.1	86.3	48.1	65.5	75.7
(b) Disable VAE γ correlation	✓		✓	-	-	-	75.8	86.8	85.1	48.5	65.9	75.0
(c) Disable Variational Inference	✓			-	-	-	75.1	85.1	84.0	48.0	64.9	74.7
(d) Enable Proposal	-	-	-		✓		75.2	86.8	86.9	46.3	65.8	74.0
(e) Enable Auxiliary and Proposal	-	-	-	✓	✓		75.8	87.2	87.1	47.0	66.6	74.3
(f) Enable Proposal and Augment	-	-	-		✓	✓	77.1	87.9	87.1	49.1	67.0	76.2
(g) Proposed Method	✓	✓	✓	✓	✓	✓	77.6	88.8	87.7	49.7	67.6	77.0

[1] In *Component-wise* section, we maintain all three branches, but disable major components of the proposed method one at a time to examine the individual contribution of each method component to the overall performance.
[2] In *Branch-wise* section, we enable all the method components, and then evaluate the major combinations of the proposed three branches.

sub-figure presents one challenge case. TrackR-CNN is a top-performing method, but still exhibits some systematic defects on highly overlapping instances, suggesting that the existing methods are challenged by the fundamental difficulty of instance segmentation tracking. Our visualization depicts comparably superior instance segmentation masks and correctly-maintained instance identities in the cluttered scenes (e.g., among multiple occluded vehicles in Figure 3 (a) and with partial occlusions caused by accessories and body parts of other pedestrians in Figure 3 (b)).

4.3. Ablation Study

We run a number of ablation experiments and report the results in Table 2.

Component-wise. Firstly, we investigate the effectiveness of main method components. Table 2 (a) shows the result of removing skip connections. We employ the skip connection scheme to supplement VAE such that the network could propagate context information to higher resolution layers in a contracting path and enables precise localization in a symmetric expanding path. After removing this scheme, the overall sMOTSA drops more than 2% in cars category and 1.6% in pedestrians. It shows that the skip connections actually help for information preservation. Table 2 (b) reports the result of removing correlation modeling for VAE latent variables. In this setting, the framework becomes a conventional VAE, in which every latent variable is assumed to be independent and identically distributed *(iid)*. It is observed that conforming strong but unrealistic *iid* assumption in standard VAE results in 1.8% sMOTSA loss in cars category. Table 2 (c) shows the result of removing VAE architecture. Thus, the model becomes an encoder-decoder framework without variational inference and correlation modeling. We can observe a notable loss in sMOTSA. This suggests that VAE architecture actually learns crucial information for VIST task.

Branch-wise. We evaluate the importance of different branches and report results in Table 2 (d)(e)(f). We use the bounding box predictions generated in the proposal branch to form object tracklets. To be able to evaluate VIST performance, we ablate the other two branches step by step for a controlled evaluation within our framework. We draw several conclusions from the results: (1) We add the auxiliary branch, which is served as a regularization to prevent the network learning to ignore latent space, and also to give significant control over the VAE to preserve meaningful semantic information. The branch is not directly involved in VIST task; however, we find this branch is effective and influences the ability of the proposed network. (2) Compared with the proposal-only model, the model with proposal and augment branch benefits from better localized bounding boxes and more discriminative features fused from difference sources, thus yields considerable improvement in sMOTSA. However, without additional supervision by the auxiliary branch, the performance is not on-par with the integrated framework. (3) With the integration of all three branches, the framework is capable of learning the embedding spatial interdependence and motion continuity in video data and brings the full potential of the proposed method.

5. Conclusion

We have introduced a unified variational autoencoder modification for reliably segmenting and tracking multiple instances in diverse and unstructured videos, where extensive object occlusions and deformations exist and affect the way by which many heretofore existing methods perform. Experiments on several distinct datasets demonstrate the superiority of the proposed method when compared to the state-of-the-art methods that adapt Mask R-CNN[18] by adding an association module to track objects over time.

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Anton Andriyenko, Konrad Schindler, and Stefan Roth. Discrete-continuous optimization for multi-target tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1926–1933. IEEE, 2012. 1, 2

[2] S Avinash Ramakanth and R Venkatesh Babu. Seamseg: Video object segmentation using patch seams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 376–383, 2014. 3

[3] Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. Semi-supervised video segmentation using tree structured graphical models. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2751–2764, 2013. 3

[4] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225. IEEE, 2014. 1, 2

[5] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2

[6] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 6, 7

[7] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer, 2010. 3

[8] Ignas Budvytis, Vijay Badrinarayanan, and Roberto Cipolla. Mot-mixture of trees probabilistic graphical model for video segmentation. In *BMVC*, volume 1, page 7. Citeseer, 2012. 3

[9] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR 2017*. IEEE, 2017. 2, 3

[10] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. 2

[11] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1189–1198, 2018. 3

[12] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. *arXiv preprint arXiv:1806.02323*, 2018. 2

[13] Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Visual tracking using attention-modulated disintegration and integration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4321–4330, 2016. 2

[14] Anna Ellis and James Ferryman. Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 135–142. IEEE, 2010. 2

[15] Qingnan Fan, Fan Zhong, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Jumpcut: non-successive mask transfer and interpolation for video cutout. *ACM Trans. Graph.*, 34(6):195–1, 2015. 3

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[17] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 5

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2, 6, 7, 8

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[20] Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon. Online multi-object tracking via structural constraint event aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1392–1400, 2016. 2

[21] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. MaskRNN: Instance level video object segmentation. In *Advances in Neural Information Processing Systems*, pages 325–334, 2017. 2, 3, 5

[22] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–802, 2018. 2, 3

[23] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018. 3

[24] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019. 2

[25] Suyog Dutt Jain and Kristen Grauman. Supervoxel-consistent foreground propagation in video. In *European conference on computer vision*, pages 656–671. Springer, 2014. 3

[26] Nikhil Ketkar. Introduction to pytorch. In *Deep learning with python*, pages 195–208. Springer, 2017. 6

[27] Margret Keuper, Siyu Tang, Yu Zhongjie, Bjoern Andres, Thomas Brox, and Bernt Schiele. A multi-cut formulation

for joint segmentation and tracking of multiple objects. *arXiv preprint arXiv:1607.06317*, 2016. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4, 5

[30] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016. 2

[31] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[32] Cheng-Hao Kuo and Ram Nevatia. How does person identity recognition help multi-person tracking? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1217–1224. IEEE, 2011. 2

[33] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 3

[34] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2

[35] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 1, 2

[36] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 2

[37] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 2

[38] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 2

[39] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6526–6535, 2018. 2, 3

[40] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 207–223, 2018. 2, 3

[41] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 90–105, 2018. 3

[42] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2953–2960. IEEE, 2009. 2

[43] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017. 2

[44] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[45] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 2

[46] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018. 2

[47] Nicolas Märki, Federico Perazzi, Oliver Wang, and Alexander Sorkine-Hornung. Bilateral space video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 743–751, 2016. 3

[48] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 6

[49] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011. 2

[50] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2

[51] Xiaofeng Ren and Jitendra Malik. Tracking as repeated figure/ground segmentation. In *CVPR*, volume 1, page 7. Citeseer, 2007. 3

[52] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[53] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2167–2176, 2017. 3

[54] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821. IEEE, 2012. 1, 2

[55] Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal video object segmentation from persistence of occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4268–4276, 2015. 3

[56] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3394, 2017. 3

[57] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M Rehg. Motion coherent tracking using multi-label mrf optimization. *International journal of computer vision*, 100(2):190–202, 2012. 3

[58] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2813, 2017. 2

[59] 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. 1, 2, 3, 6, 7

[60] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 2, 3

[61] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association with online target-specific metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241, 2014. 2

[62] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4854–4863, 2018. 2

[63] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4705–4713, 2015. 1, 2

[64] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018. 2

[65] Bo Yang and Ram Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1918–1925. IEEE, 2012. 1, 2

[66] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 3, 6, 7

[67] Hongkai Yu, Youjie Zhou, Jeff Simmons, Craig P Przybyla, Yuewei Lin, Xiaochuan Fan, Yang Mi, and Song Wang. Groupwise tracking of crowded similar-appearance targets from low-continuity image sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 952–960, 2016. 2

[68] Xuemei Zhao, Dian Gong, and Gérard Medioni. Tracking using motion patterns for very crowded scenes. In *Computer Vision–ECCV 2012*, pages 315–328. Springer, 2012. 1, 2

[69] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 2

[70] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Unsupervised online video object segmentation with motion property understanding. *IEEE Transactions on Image Processing*, 29:237–249, 2019. 2, 3