Neurocomputing 655 (2025) 131439

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# A memory-based conditional neural process for video instance segmentation

Kunhao Yuan [a], Gerald Schaefer [b], Yu-Kun Lai [c], Xiyao Liu [d], Lin Guan [b], Hui Fang [b,*]

[a] Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK
[b] Department of Computer Science, Loughborough University, Loughborough, UK
[c] School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[d] School of Computer Science and Engineering, Central South University, Changsha, China

## HIGHLIGHTS

- A memory-based conditional neural process.
- Reliability modelling for object detection.
- Uncertainty-based dynamic training sample selection.
- Contrastive instance tracking.

## ARTICLE INFO

Communicated by G. Hu

*Keywords:*
Video instance segmentation
Memory network
Neural processes
Uncertainty

## ABSTRACT

Video instance segmentation (VIS) is an evolving research topic in computer vision that aims to simultaneously detect, segment, and track semantic objects across multiple video frames. However, existing VIS methods are typically unaware of the reliability of the training samples from insufficient and imbalanced datasets, leading to suboptimal performance. To address this challenge, we propose a memory-based conditional neural process (MemCNP) module to exploit the strengths of both memory networks and the CNP model which handles heterogeneous latent space distributions for reliable modelling with insufficient data. Our MemCNP utilises predicted uncertainty to regularise VIS predictions as well as to identify reliable samples for effective training. Notably, our MemCNP is model-agnostic and can thus be seamlessly integrated into various VIS models to improve their performance. Extensive experiments on the YouTube-VIS and OVIS datasets demonstrate the effectiveness of MemCNP regardless of the underlying model architecture.

## 1. Introduction

Video instance segmentation (VIS), first introduced in YouTube-VIS 2019 [1], aims to simultaneously detect, segment and track objects in video sequences, and is useful in various applications such as instance-level video editing, autonomous driving and augmented reality. In contrast to conventional image-level object detection and image segmentation [2], VIS requires a holistic approach to consider temporal dynamics and spatial relationships between objects as well as fine-grained parsing of videos.

Many existing VIS methods extend conventional image-level instance segmentation algorithms such as Mask R-CNN [2] to process videos frame-by-frame [1,3,4]. However, they are unable to fully exploit the temporal contexts in video sequences [5]. While some other methods, such as [6–8], leverage temporal cues to improve the accuracy and robustness of VIS, they are computationally expensive and infeasible for real-time applications. To strike a balance between efficiency and accuracy, semi-online methods [9,10] process down-sampled or short video clips. Recent trends in VIS focus on holistic video-level instance association [11], context-aware tracking [12], temporal feature propagation [13], and object re-identification during rapid frame changes [14], further distinguishing VIS from image-based instance segmentation.

An overlooked issue in current VIS approaches is the lack of knowledge about the reliability of the extracted regions of interest (RoIs) during the training stage, which renders the selection of training candidates non-trivial. Existing methods rely on RoI overlap [3,4] or temporal

* Corresponding author.
*Email address:* h.fang@lboro.ac.uk (H. Fang).

https://doi.org/10.1016/j.neucom.2025.131439
Received 10 March 2025; Received in revised form 20 August 2025; Accepted 30 August 2025
Available online 1 September 2025

differences between frames [5,15] for model updates which can become imprecise and inflexible and lead to performance degradation. Similarly, frames randomly sampled from a video sequence may not support reliable feature extraction. Since many frames contain motion blurred objects, occlusions, and rapid dis- and re-appearances of objects, the use of these RoIs can compromise the overall model performance. Typical two-stage methods [1,2,5] use region proposal networks (RPNs) [2] to first eliminate unreliable candidates and subsequently perform training on the remaining candidates. Although this approach improves reliability, it introduces a non-end-to-end paradigm, sacrificing efficiency and ease of post-processing. In contrast, one-stage methods [3,16] require prior knowledge of data distributions to manually establish hyper-parameters that can efficiently select reliable candidates against an overwhelming number of false positives. Recent DETR-based methods [4,6,9] replace feature pyramid networks (FPNs) with transformers and leverage Hungarian matching for candidate selection, but suffer from slow convergence and a high demand for training data, limiting their practical applicability. In addition, all these methods assume a well-structured underlying data distribution, which is not always the case when dealing with insufficient video sequences.

Inspired by the recent conditional neural process (CNP) model [17], in this paper, we present a memory-based CNP (MemCNP) to tackle the above research challenges and further improve VIS model reliability. Our proposed MemCNP learns dataset-scale contexts from a memory bank which is used to ensure the outputs from the VIS model are statistically meaningful, thus enhancing model reliability and simplifying training candidate selection. We demonstrate, by explicitly modelling bounding box regressions as functions of distributions, that the learned variance of each distribution can serve not only as an estimation of detection uncertainty but also as a criterion for confident candidate selection.

Our contributions in this paper are:

- We propose MemCNP to improve VIS model reliability. The MemCNP is the first approach to exploit the CNP model to ensure that the outputs from the trained VIS models are more explainable and reliable from a statistical perspective. In addition to enhancing the representation used for bounding box regression in VIS, we also design a CNP loss along with uncertainty minimisation to explicitly measure uncertainties of predictions.
- We introduce uncertainty-based candidate selection which overcomes the limitations of heuristically designed selection, enabling dynamic exploitation of high-quality samples for multi-task learning in VIS.
- Our method is model-agnostic and can be easily applied to existing VIS methods. Extensive experiments on the YouTube-VIS [18] and OVIS datasets [5] demonstrate its effectiveness regardless of the VIS backbone architecture.

Our code is available at https://github.com/SCoulY/MemCNP-VIS.

The remainder of the paper is organised as follows. Related work on video instance segmentation, neural processes and memory networks is reviewed in Section 2. Our proposed approach is then explained in detail in Section 3. Experimental results are reported and discussed in Section 4, and Section 5 concludes the paper.

## 2. Related work

### 2.1. Video instance segmentation

Video instance segmentation is initially proposed in [1] as an advancement of image instance segmentation (IIS) [2] and video object segmentation (VOS) [19]. Restricted by computational resources, early methods extend existing image instance segmentation models with a tracking method to process videos frame-by-frame in an online manner [18]. To improve VIS efficiency, FCOS [16] replaces the instance-wise region proposal network with pixel-wise dense convolutions, enabling training in a single phase. Extending FCOS [16], in [3,20] instance-independent coefficients are incorporated into the segmentation branch to achieve improved performance. To handle challenging scenarios with highly-occluded objects, [5] introduces a temporal calibration module between key-reference frames to identify these occlusions. To further utilise temporal correlations, 3D convolutional models [10] and transformers [8] are exploited to process video sequences as a whole, and are capable of enhancing feature representations extracted from the full spatio-temporal information in videos. Furthermore, TriANet [21] proposes a holistic attention-module to exploit spatial-, temporal-, and channel-wise contexts for video segmentation. In SeqFormer [6] and IDOL [4], the concept of learnable queries is introduced, wherein these queries are iteratively associated with object instances, resulting in a simplification of post-processing stages and enhanced video instance segmentation performance. Subsequent approaches concentrate on integrating the training and inference phases, employing methods such as unified label assignment [11] or the implementation of a consistent memory bank [22], thereby facilitating more stable instance discrimination. Other recent advancements address challenges in inconsistent object tracking [12], foundation model-based enhancement [23] and re-identification [14] that occur due to rapid changes between video frames.

While recent methods like IDOL [4], GenVIS [11] and DVIS [14] significantly advance the field by refining VIS architectures or unifying training strategies, our work introduces a complementary approach with focus on the overlooked problem of prediction reliability. Our proposed MemCNP is a model-agnostic module that leverages probabilistic modelling to estimate prediction uncertainty. This allows it to enhance existing VIS frameworks, including convolution-based [2], DETR-based [24] and MaskFormer-based [25] ones—by reliable sampling and improved tracking. This focus on being a versatile, reliability-enhancing plug-in is a key distinction from architecture-specific solutions.

### 2.2. Probabilistic models

Probabilistic modelling is a well-researched avenue to interpret model behaviour and improve model reliability such as Bayesian neural networks and variational inference [26–28]. Minimising predictive variance of model outputs is an effective strategy to improve regression and segmentation performance [29], while, sharing a similar idea, non-local probabilistic loss functions can be used to build reliable object detectors [30,31]. However, they yield only an insignificant performance boost over conventional models when non-probabilistic metrics, such as mean average precision (mAP), are used for evaluation.

Neural processes (NPs) offer an alternative approach to probabilistic modelling [32]. Instead of building probabilistic distributions directly from model outputs, example-based contexts are leveraged to generate outputs using functions over distributions, enabling the estimation of uncertainty in prediction. Recent NP advancements focus on improving either computational efficiency through conditional neural processes (CNPs) [17] or prediction accuracy through attention mechanisms [33]. However, an NP is exclusively designed for learning from a dataset of small sample size and requires a forward pass of the entire context set, thus hindering its generalisation to larger datasets.

Unlike Bayesian and variational methods, our MemCNP models function-level (bounding box-level) uncertainty by learning a distribution over functions conditioned on prototypical context, yielding semantically grounded, data-driven predictions. Additionally, the incorporation of a learned memory bank provides dataset-scale contextualisation, addressing the generalisation limitations of NPs. Beyond estimation, MemCNP actively uses uncertainty during training, using uncertainty-based sample selection and contrastive tracking, turning

it into a learning signal, a feature absent in standard Bayesian or variational approaches.

### 2.3. Memory networks

Recurrent neural networks (RNNs), such as long short-term memory (LSTM) [34] and gated recurrent unit (GRU) [35] models, have been primary means to implement memorisation in neural networks. With the help of hidden units, RNNs can memorise previous inputs to enhance the representation at a new timestamp. To overcome the limited accessibility of RNNs, memory networks were proposed in [36] to utilise contexts from individual items. Recently, memory networks have also gained popularity in various computer vision tasks such as captioning [37], deraining [38], video object segmentation [39] and video instance segmentation [22]. These approaches apply iterative clustering to group inputs into representative prototypes to enhance feature representations even though they are arbitrarily distributed in the latent space. By interacting with the prototypes, memory networks demonstrate robustness in identifying unseen instances, leading to improved performance compared to approaches without explicit memories.

Distinguishing it from other memory networks, our proposed MemCNP integrates both feature representation and the context of the ground truth (*i.e.*, the bounding box in VIS) into each memory item. This explicit correlation between feature representation and regression target significantly improves the performance of VIS.

## 3. Proposed approach

Our proposed MemCNP model serves as the core module to enhance VIS by incorporating probabilistic modelling and memory networks. MemCNP not only improves bounding box regression by learning dataset-scale contexts but also generates uncertainty estimates, which play a crucial role in refining training sample selection and instance tracking. Specifically, we leverage these uncertainty estimates in uncertainty sample selection (USS) to dynamically identify high-confidence samples for robust multi-task learning. Additionally, we introduce contrastive instance tracking, which benefits from the uncertainty-aware feature representations produced by MemCNP, allowing for improved instance association across frames. Fig. 1 provides an overview of our framework, which seamlessly integrates these components to enhance VIS reliability and performance. In the following, we describe each component in detail.

### 3.1. VIS pipeline

The inputs to online VIS framework include a pair of image frames, the key frame $\mathrm{F}_{key}$ and the reference frame $\mathrm{F}_{ref}$ (both $\in \mathbb{R}^{3\times H\times W}$, where $H$ and $W$ are the height and width of image frames). These frames are randomly selected from the same video clip, ensuring a sampling within a short time interval. After processing by a backbone, such as a convolutional neural network (CNN) [2] or a vision transformer (ViT) [40,41], the features from individual stages are merged by a neck network into a feature pyramid.

For fully convolutional methods, VIS is then performed on the dense feature representation $Z \in \mathbb{R}^{C\times H\times W}$, where $C$ is the number of feature channels. The first step here is dense object detection, where independent left, top, right, and bottom distances ($l$, $t$, $r$, $b$) are regressed from each pixel location, representing the four distances to the nearest object bounding box edges. Subsequently, confident candidates are chosen for segmentation, classification and tracking. In particular, segmentation is achieved by classifying the pixels within the detected bounding boxes into foreground and background. Object classification is then accomplished by categorising the detected objects into $K$ pre-defined classes. Finally, the central vector representations of the detected objects in the key frame are extracted for comparison with target representations from the reference frame to achieve tracking.

The loss function for VIS is typically formulated based on valid candidates and incorporates losses for bounding box regression, classification, segmentation, and tracking as

$$\mathcal{L}_{vis} = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{track}\mathcal{L}_{track}, \tag{1}$$

where $\mathcal{L}_{bbox}$ is an intersection-over-union (IoU)-oriented loss, $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$ are binary cross-entropy losses, $\mathcal{L}_{track}$ is typically an $n$-way cross-entropy loss depending on the number of objects present in the reference frame, and the $\lambda$ parameters are used to balance the loss terms.

It should be noted that our proposed method is not limited to online and CNN-based frameworks but is equally applicable to offline and transformer-based methods, as we will also demonstrate in Section 4.

### 3.2. Memory-based conditional neural process

A conditional neural process [32] is a neural network model that learns a distribution $Q_\theta$ over functions $f(x)$ to predict targets $x's$ conditioned on a fixed observation set $O$. In particular, the training set of a total of $U$ samples is divided into two subsets: the observed set denoted

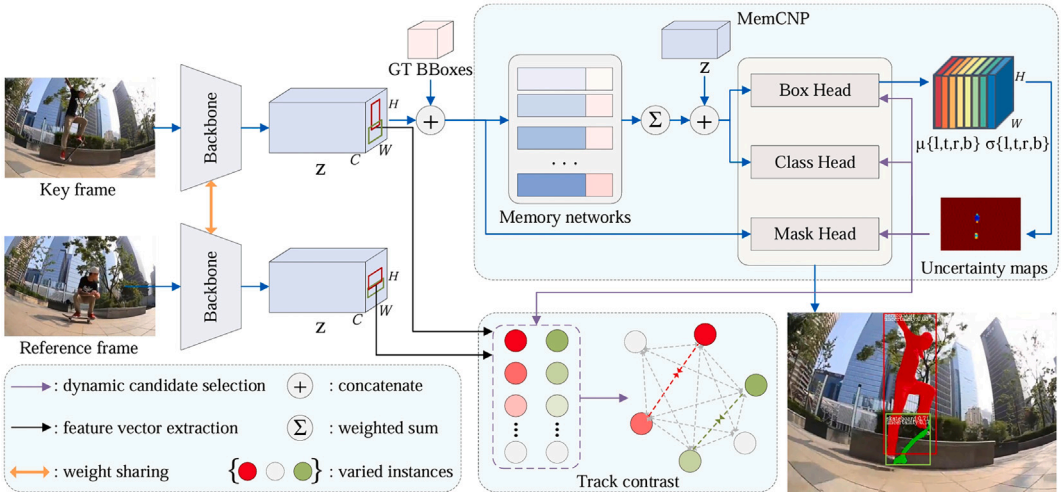


**Fig. 1.** The pipeline of our proposed method, highlighting the probabilistic MemCNP module that facilitates dynamic candidates selection for box, class and mask heads via estimated uncertainty maps. Individual memory item is composed of the feature vector and corresponding context from ground truths (bounding boxes). Shapes that possess the same colour denote the same instance across different frames and are considered positive pairs in the contrastive tracking, while the rest are treated negative.

as $O$, comprising $N$ pairs of observed samples and their corresponding labels $(x_n, y_n)_{n=1}^{N}$, and the target set denoted as $T$, which contains the remaining pairs $(x_i, y_i)_{i=N+1}^{U}$.

First, an encoder $E_\theta$ maps the sample pairs from the observed set to compact latent representations $r_n \in \mathbb{R}^d$, *i.e.*, $r_n = E_\theta(x_n, y_n)$. After aggregating all latent representations $r = \frac{1}{N}\sum_{n=1}^{N} r_n$, a decoder $D_\theta$ is applied to map a target input $x_i$ of the pair $(x_i, y_i)$ drawn from $T$, to the output conditioned on the aggregated context representation $r$, which can be expressed as

$$\begin{aligned} Q_\theta(f(T) \mid O, T) &= \prod_{(x,y)\in T} Q_\theta(f(x) \mid O, x) \\ &= \prod_{(x,y)\in T} Q_\theta(D_\theta\left(x \mid E_\theta(O)\right)) \\ &= \prod_{i=N+1}^{U} Q_\theta\left(D_\theta\left(x_i \mid r\right)\right). \end{aligned} \tag{2}$$

For regression, the network $\theta$ is trained to parameterise two statistics, $\mu_i$ and $\sigma_i$, to define a Gaussian distribution such that $Q\left(f\left(x_i\right) \mid O, x_i\right) \sim \mathcal{N}\left(\mu_i, \sigma_i\right)$ by minimising the negative conditional log-likelihood through

$$\begin{aligned} \mathcal{L}(\theta) &= -\mathbb{E}_{\{x_i,y_i\}_{i=N+1}^{U}}\left[\mathbb{E}_N\left[\log Q\left(f\left(x_i\right) \mid O_N, x_i\right)\right]\right] \\ &= -\mathbb{E}_{\{x_i,y_i\}_{i=N+1}^{U}}\left[\mathbb{E}_N\left[\log Q\left(y_i \mid O_N, x_i\right)\right]\right]. \end{aligned} \tag{3}$$

CNP, whose processing pipeline is illustrated in Fig. 2, and its variants [17,32,33] have shown advantages over non-probabilistic methods in tasks such as classification, regression, and image completion, especially when trained on small datasets.

VIS is different from other applications with insufficient and imbalanced datasets. On one hand, video sequences comprise a large amount of training data. On the other hand, the diversity and quality of the training samples render building a reliable DNN model a challenging task. To exploit CNPs in such a scenario, we propose a CNP model that is supported by a memory bank to efficiently exploit contexts extracted from a large observation set through memory units during each training iteration. Our memory module is designed to contain the most typical representations extracted from the encoder, which can be a CNN [2] or a Transformer model [40], functioning as references and conditioning in a probabilistic model. Thus, our MemCNP is used to retain features within the latent space while being independent of the underlying encoder architecture.

The processing pipeline of our proposed MemCNP is illustrated in Fig. 3. We initialise a memory bank $M \in \mathbb{R}^{k\times(d+4)}$, storing $k$ items with $d+4$ dimensions, where the 4 extra dimensions encompass the left, top, right, and bottom distances ($l$, $t$, $r$, $b$) to the nearest ground truth bounding box (if available).

While in CNP $x$ is directly concatenated with its corresponding regressed value $y$ before passing into the encoder, in the image domain, low-level pixel values $x$ do not correspond to meaningful semantic classes or bounding boxes and thus the targets $y \in \mathbb{R}^4$. Therefore, to construct relevant input-target pairs, we first encode an image $\mathrm{F} \in \mathbb{R}^{H\times W\times 3}$

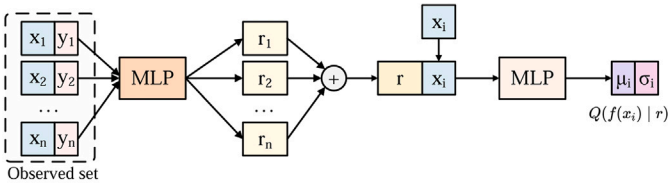


**Fig. 2.** Illustration of the CNP pipeline. The observed set is predefined, with both input features $x_n$ and corresponding outputs $y_n$ bounded to construct the dataset-scale context, serving as a condition for unseen samples $x_i$.

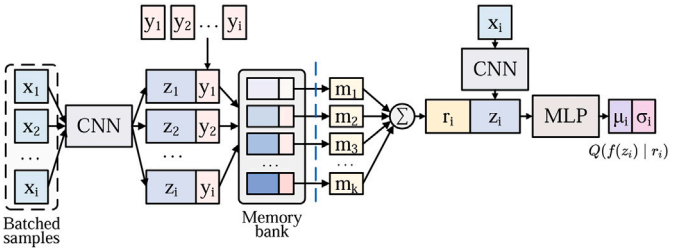


**Fig. 3.** Illustration of MemCNP pipeline. The batched samples aren't directly associated with their respective outputs, such as bounding box coordinates. Instead, they undergo processing to derive a globally distinctive intermediate representation $z_i$. Subsequently, this representation is combined with the most relevant bounding box coordinates $y_i$ from memory, enhancing the comprehensive representation of $z_i$ within the current batch. On the left of the dashed vertical line the operations are exclusive to the training phase, while the operations on the right are those shared between training and inference.

into dense latent representations $Z \in \mathbb{R}^{H\times W\times d}$ and then compute (dot product) similarities between latent representations and memory items as $S = ZM^T$. The most relevant memory item $m_{\hat{j}}$ for a single latent vector $z_i$ is identified by

$$\hat{j} = \arg\max_j S_{ij}. \tag{4}$$

After encoding, the latent vector becomes globally diversified, allowing it to distinguish objects of different semantic classes and geometric appearances. We therefore concatenate each latent $z_i \in \mathbb{R}^d$ with the regression target, *i.e.* with the distances $y_i \in \mathbb{R}^4$ from its location in the feature map to the four edges of the nearest ground truth bounding box, to construct the paired query item $\tilde{z}_i \in \mathbb{R}^{d+4}$. The most relevant memory item is then updated as

$$m_{\hat{j}} \leftarrow \tau m_{\hat{j}} + (1-\tau)\tilde{z}_i, \tag{5}$$

where $\tau \in [0, 1]$ is the moving average updating rate.

We adopt a moving average update rather than hard replacement or attention-based updates to ensure temporal stability and robustness. This smooths memory evolution over time, mitigating abrupt changes and reducing sensitivity to noisy inputs, while also offering computational efficiency, making it well-suited for dense VIS tasks.

In CNP, the rich context $r$ from the entire observed set is aggregated by averaging individual latent representations. However, $r$ is shared for all target inputs $x_i$, making downstream features less representative. To enhance the distinctiveness of the context, we instead use a weighted average on memory items through soft-attentive scores, calculated as

$$a_{ij} = \frac{\exp\left(s_{ij}\right)}{\sum_{j=1}^{k}\exp\left(s_{ij}\right)}, \tag{6}$$

where $i$ and $j$ index the target samples and memory items, respectively. The input independent context is then obtained as

$$r_i = \sum_{j=1}^{k} a_{ij} m_j. \tag{7}$$

Similar to the regression objective in CNP, in our MemCNP we formulate a conditional Gaussian distribution $Q\left(f\left(z_i\right) \mid r_i\right) \sim \mathcal{N}\left(\mu_i, \sigma_i\right)$ by minimising the pixel-wise negative log-likelihood as

$$\begin{aligned} \mathcal{L}_{nll} &= -\mathbb{E}_Z\left[\mathbb{1}\left(\min\left(y_i\right) > 0\right)\log Q\left(f\left(z_i\right) \mid r_i\right)\right] \\ &= -\mathbb{E}_Z\left[\mathbb{1}\left(\min\left(y_i\right) > 0\right)\log Q\left(\hat{y}_i \mid r_i\right)\right], \end{aligned} \tag{8}$$

where $\mathbb{1}$ is an indicator function used to filter out any input locations within a valid ground truth bounding box, and $\hat{y}_i$ represents the

ground truth $y_i$ in normalised coordinate format. The benefits of explicitly modelling the regression target as a Gaussian variable are two-fold. First, the probabilistic loss function works as an auxiliary to the standard geometric $L_1$ loss or IoU loss, making the prediction more interpretable from a probability perspective. Second, apart from outputting the mean value $\mu_i$ for each target, the network also derives the predictive uncertainty in the form of the variance $\sigma_i$. This allows for self-evaluation of the network's predictions, and thus an assessment of its reliability.

To make our model more robust against local perturbations in individual predictions, we replace the unitary Gaussian distribution $Q\left(f\left(z_i\right) \mid r_i\right)$ in Eq. (8) with a Gaussian mixture model $Q(b_j)$. Assuming that for ground truth $b_j$ there are $I$ associated candidate pixels, we obtain this as

$$Q(b_j) \sim \sum_{i=1}^{I} \phi_i \mathcal{N}(\mu_i, \sigma_i), \tag{9}$$

where $\phi_i$ is the mixing weight defined as the centredness of location $i$ regarding box $b_j$. The overall loss for all ground truths is then calculated as

$$\mathcal{L}_{nll} = -\frac{1}{J} \sum_{j=1}^{J} \log(Q(b = b_j)), \tag{10}$$

where $J$ is the number of ground truths.

Since the variances of the bounding box predictions are less meaningful without any constraints, we enforce a regularisation by incorporating classification probabilities and the intersection over union (IoU) between the regressed box and the ground truth box. For a ground truth box $b_j$ and a corresponding paired candidate box $\tilde{b}_j^i$, we estimate an approximate uncertainty as

$$u_j^i = \left(1 - \tilde{P}_j^i\right) + \left(1 - \mathrm{IoU}\left(\tilde{b}_j^i, b_j\right)\right), \tag{11}$$

where $\tilde{P}_j^i$ represents the softmax target class probability of the ground truth box. With this uncertainty taking both classification and regression errors into consideration, we can construct a regularisation term on the free-form predictive uncertainty as

$$\mathrm{R}_\sigma = \frac{1}{JI} \sum_{j=1}^{J} \sum_{i=1}^{I} \|\sigma_j^i - u_j^i\|_p, \tag{12}$$

where $\|\cdot\|_p$ denotes the $p$-norm. This term can both prevent the unrestricted expansion of free-form uncertainty, providing reasonable bandwidth for maximising the log-probability in $\mathcal{L}_{nll}$, and enhance the distinctiveness among candidate boxes by constraining $\sigma_j^i$ with individual uncertainties rather than solely relying on the population-level constraint $\mathcal{L}_{nll}$.

These steps collectively ensure that the predicted uncertainties are both meaningful and bounded. Our approach begins by framing bounding box regression probabilistically through a negative log-likelihood loss (Eq. 8), encouraging the model to learn both a mean and variance for its predictions. To enhance robustness, we then model the multiple candidates associated with a single ground truth box as a Gaussian Mixture Model (GMM) (Eq. 9), which captures local variability and reduces reliance on any single candidate. The final loss (Eq. 10) aggregates these more reliable GMM-based predictions. Finally, to ground the learned variances in task-specific performance, we introduce an auxiliary supervision signal (Eqs. 11 and 12) that uses classification confidence and IoU as a proxy for uncertainty. This constraint regularises the predicted variance, preventing it from growing arbitrarily while aligning it with tangible reliability cues.

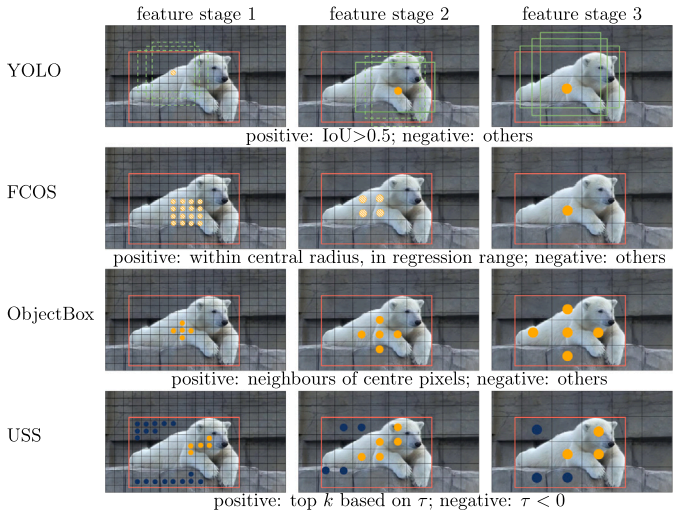


**Fig. 4.** Different sample selection methods in comparison to our proposed USS strategy. Solid yellow dots indicate positive candidates that satisfy the selection criterion, dashed candidate boxes and striped dots represent candidates that do not satisfy it, blue dots represent our trusted negative candidates in USS.

Finally, we obtain the loss function of our MemCNP module as

$$\mathcal{L}_{cnp} = \mathcal{L}_{nll} + \mathrm{R}_\sigma, \tag{13}$$

and the $\mathcal{L}_{bbox}$ in Eq. (1) becomes

$$\mathcal{L}_{bbox} = \mathcal{L}_{GIoU} + \alpha \mathcal{L}_{cnp}, \tag{14}$$

where $\alpha$ is used to balance the probabilistic and non-probabilistic regressions.

### 3.3. Uncertainty sample selection

As discussed in Section 2.1, the detected RoIs are subsequently used for segmentation, tracking and classification [1,2]. To identify reliable samples for the multi-task learning of VIS, we introduce a novel sample selection strategy named uncertainty sample selection (USS). This is built on top of the predictive uncertainties and their corresponding centredness values to prioritise samples that are likely to yield accurate results. Given a ground truth box $b_j$, a valid candidate set $I$ contains all pixels located within the box. For every pixel in $I$, we calculate the confidence score $\tau$ by combining the predictive uncertainty and the centredness score as

$$\tau_j^i = -\lambda_\tau \sigma_j^i + (1 - \lambda_\tau) c_j^i, \tag{15}$$

where $c_j^i$ is the centredness of location $i$ regarding ground truth box $b_j$, and $\lambda_\tau$ allows for balancing the two terms when the predictive uncertainty is unstable at the beginning of training.

In order to perform multi-scale regression, for every feature stage, we select the $k$ samples with largest $\tau$ as positive samples, while samples with $\tau < 0$ are selected as negative samples, as illustrated in Fig. 4. It should be noted that this is different from existing methods [3,16,42], where a pixel inside the box is marked either as positive or negative (see rows 1–3 in Fig. 4). In contrast, our method is capable of identifying both confident positive and negative candidates. In addition, USS operates dynamically and uniformly on feature stages with different scales, removing the manual design of anchor boxes [43,44], heuristic constraints [16], and off-centre object placements [42] (such as the bottom-most sample in feature stage 3 for ObjectBox in Fig. 4) associated with other candidate selection methods.

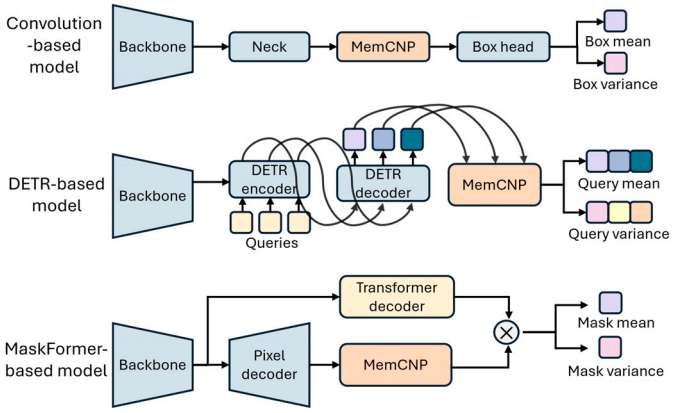


**Fig. 5.** Illustration of model-agnostic feature of the MemCNP module.

### 3.4. Contrastive instance tracking

Online VIS methods use key/reference frame pairs for object tracking in video sequences, matching the feature vector of the target instance in the key frame to one of the feature vectors related to objects in the reference frame or an unmatched dummy instance if the target is absent [5,16,18]. However, this strategy only compares the target feature vector to those objects that appear in the reference frame, ignoring the fact that the target instance can be confused with other instances in either the key frame or the reference frame.

Inspired by the recent success of contrastive learning in classification, detection, and segmentation tasks [45,46], we propose a contrastive instance tracking approach that leverages information from diverse instances both within and across frames, distinguishing it from IDOL [4]. Furthermore, unlike IDOL, which selects positive and negative samples based on bipartite matching costs, our method relies on uncertainty-guided sample selection, enabling more informed and adaptive contrastive learning.

Let $z_i$ and $\tilde{z}_i$ be the tracking feature vectors extracted from the centre of a candidate box in the key frame and the corresponding ground truth box in the reference frame, respectively, and $z_{j\neq i}$ be a tracking vector from the remaining candidate boxes related to different ground truths selected by USS in a training batch. Our tracking process brings positive pairs ($z_i$ and $\tilde{z}_i$) closer while simultaneously pushing negative pairs ($z_i$ and $z_j$) farther apart as

$$\mathcal{L}_{track} = -\log\left(\frac{\sum_{Z^+}\exp(z_i\cdot\tilde{z}_i/\gamma)}{\sum_{Z^+}\exp(z_i\cdot\tilde{z}_i/\gamma)+\sum_{Z^-}\exp(z_i\cdot z_j/\gamma)}\right), \tag{16}$$

where $Z^+$ and $Z^-$ are the positive and negative sample pairs in a batch, $\cdot$ represents the dot product, and $\gamma$ is a temperature parameter (set to 0.1 following [45–47]).

## 4. Experimental results

### 4.1. Experimental setup

We conduct our experiments on the YouTube-VIS 2021 [18] and OVIS datasets [5]. YouTube-VIS comprises 3859 video clips, 8171 unique video instances and 232 k instance annotations spanning across 40 categories, while OVIS is a more challenging dataset with severe occlusions and contains 901 videos, 5223 unique instances and 296 k instance masks over 25 categories. Notably, sequences in OVIS contain more object instances and longer video clips, thus demanding more computational resources.

We mainly build upon the well-established fully convolutional instance segmentation method from Ref. [3] as the foundation of our algorithm, which allows us to compare different backbones and neck network architectures effectively. To further demonstrate the model-agnostic nature of our approach and evaluate its applicability across diverse VIS paradigms, we also integrate MemCNP into typical DETR-based offline and online frameworks, including SeqFormer [6], IDOL [4], and the MaskFormer DVIS-DAQ [14]. Fig. 5 illustrates the integration of MemCNP into the various frameworks, while block-wise model parameters and inference speed are listed in Table 1. As we can also see, for the classical convolution-based SipMask, our approach results in a ~7 % slowdown, whereas the impact is lower for the more advanced DETR-based methods IDOL and SeqFormer (~2 %) and the MaskFormer method DVIS-DAQ (~5 %).

Following [2,3,16], training is performed for 12 epochs by default, with the input image size varying from 360×640 to 480×960 and the loss hyper-parameters $\lambda_{bbox}, \lambda_{cls}, \lambda_{seg}$ *and* $\lambda_{track}$ set to 1. We adopt $\tau = 0.999$ and $\gamma = 0.1$ based on the best practices outlined in [38] and [46], respectively. To avoid overflow of the log-probability losses, $\alpha$ is set to $\frac{1}{\#\text{ supporting samples}}$, while configuring the balancer $\lambda_\tau$ to 0.5 for equal attention between uncertainty and centredness when performing sample selection. For our USS strategy, we select the top $k = 5$ as positive samples for probabilistic modelling and contrastive tracking). This value empirically provides the best balance between sample quality and diversity; a smaller $k = 3$ ($-1.1$ AP) risks insufficient supervision, while a larger $k = 7$ ($-0.7$ AP) may introduce noisy false positives that destabilize training.

Due to computational demands, SipMask, IDOL, and DVIS-DAQ are trained with batch sizes of 8, 4, and 2 per GPU, respectively. Training is conducted on 8 NVIDIA V100 GPUs, while inference and speed measurements are performed on a single RTX 4090 GPU. Results are obtained from the same dataset on which the model is trained, except for those marked (with † in the tables) which follow the commonly used COCO co-training strategy as in [6,9,11,14,22]. We report the standard metrics $AP_{50:95}$ (abbreviated as AP hereafter), $AP_{50}$ and $AP_{75}$ (the subscripts denote the corresponding mask IoU thresholds), and include the $AR_1$ and $AR_{10}$ results, giving the average recall in videos with at most 1 and 10 object instances. These metrics differ from standard AP and AR in object detection by using various mask IoU thresholds instead of box IoUs, and provide complementary performance assessments since they simultaneously evaluate segmentation with regard to both spatial and temporal

**Table 1**
Model parameters (in millions) and inference speed (in frames per second, FPS) measured on a single NVIDIA RTX 4090 GPU.

| | SipMask | | SeqFormer | IDOL | | DVIS-DAQ | |
|---|---|---|---|---|---|---|---|
| | ResNet-50 | ViT-b | ResNet-50 | ResNet-50 | Swin-L | ResNet-50 | ViT-L |
| Backbone | 23.5 | 85.8 | 23.2 | 23.2 | 195.2 | 23.5 | 327.5 |
| Neck | 3.9 | 3.0 | 16.4 | 11.4 | 9.7 | 20.7 | 20.8 |
| DenseHead | 6.6 | 6.6 | 9.5 | 9.4 | 9.4 | 18.9 | 19.0 |
| total | 34.0 | 95.4 | 49.1 | 44.0 | 214.4 | 63.1 | 367.3 |
| FPS w/ MemCNP | 26.8 | 20.8 | 47.3 | 29.1 | 18.0 | 35.7 | 17.6 |
| FPS w/o MemCNP | 28.8 | 22.4 | 47.6 | 29.8 | 18.3 | 37.4 | 17.7 |

**Table 2**
VIS results on YouTube-VIS 2021 validation set. Best results are **bolded**. † Indicates model is trained with COCO co-training strategy.

| Method | Backbone | Neck | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|---|---|
| SipMask | Res-50 | FPN | 27.5 | 47.8 | 27.9 | 26.4 | 31.9 |
| MaskTrack | Res-50 | FPN | 28.6 | 47.6 | 29.5 | 26.6 | 33.4 |
| STEM-Seg | Res-50 | FPN | 30.6 | 50.7 | 33.5 | 31.6 | 37.1 |
| CrossVIS | Res-50 | FPN | 34.2 | 54.4 | 37.9 | 30.4 | 38.2 |
| SeqFormer | Res-50 | DETR | 37.7 | 58.3 | 41.6 | 34.8 | 46.2 |
| EfficientVIS | Res-50 | FTSA | 37.9 | 59.7 | 43.0 | 40.3 | 46.6 |
| IDOL | Res-50 | DETR | 43.9 | 68.0 | 49.6 | 38.0 | 50.9 |
| GenVIS† | Res-50 | UVLA | 47.1 | 67.5 | 51.5 | **41.6** | 54.7 |
| DVIS-DAQ† | Res-50 | MaskFormer | 48.8 | 70.0 | 53.1 | 39.7 | 55.8 |
| Ours† | Res-50 | MaskFormer | **49.7** | **70.9** | **54.7** | 41.2 | **57.7** |
| TeViT | ViT-b | MHSA | 37.9 | 61.2 | 42.1 | 35.1 | 44.6 |
| SipMask | ViT-b | FPN | 38.1 | 57.8 | 41.4 | 33.6 | 42.6 |
| IDOL | Swin-L | DETR | 56.1 | 80.8 | 63.5 | 45.0 | 60.1 |
| GenVIS† | Swin-L | UVLA | 59.6 | 80.9 | 65.8 | **48.7** | 65.0 |
| DVIS-DAQ† | ViT-L | MaskFormer | 60.4 | 81.2 | 68.1 | 46.9 | 66.1 |
| Ours† | ViT-L | MaskFormer | **61.6** | **83.0** | **69.6** | 48.4 | **67.3** |

**Table 3**
VIS results on OVIS validation set. Best results are **bolded**. † indicates model is trained with COCO co-training strategy.

| Method | Backbone | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|---|
| SipMask | Res-50 | 10.2 | 24.7 | 7.8 | 7.9 | 15.8 |
| MaskTrack | Res-50 | 10.8 | 25.3 | 8.5 | 7.9 | 14.9 |
| STEM-Seg | Res-50 | 13.8 | 32.1 | 11.9 | 9.1 | 20.0 |
| CrossVIS | Res-50 | 14.9 | 32.7 | 12.1 | 10.3 | 19.8 |
| CMaskTrack | Res-50 | 15.4 | 33.9 | 13.1 | 9.3 | 20.0 |
| IDOL | Res-50 | 29.5 | 50.1 | 29.4 | 14.9 | 37.1 |
| GenVIS† | Res-50 | 35.8 | 60.8 | 36.2 | 16.3 | 39.6 |
| DVIS-DAQ† | Res-50 | 37.4 | 63.2 | **37.8** | 16.7 | **44.5** |
| Ours† | Res-50 | **37.7** | **64.3** | 37.7 | **16.9** | 44.3 |

dimensions. This involves summing the intersections across every frame and dividing by the union across all frames as

$$\mathrm{IoU}(i,j) = \frac{\Sigma_{t=1}^{T} \left| \mathcal{M}_t^i \cap \tilde{\mathcal{M}}_t^j \right|}{\Sigma_{t=1}^{T} \left| \mathcal{M}_t^i \cup \tilde{\mathcal{M}}_t^j \right|}, \tag{17}$$

where $\mathcal{M}_t^i$ is the predicted mask of instance $i$ at time $t$, and $\tilde{\mathcal{M}}_t^j$ denotes the corresponding ground truth mask at time $t$. Unless otherwise noted, all reported results are averages over 3 random runs.

### 4.2. Results and comparison with SOTA

We compare the performance of our proposed method with several state-of-the-art (SOTA) VIS methods and report the obtained results obtained on the YouTube-VIS 2021 validation set in Table 2. As we can see from there, when employing a standard ResNet-50 backbone, our method (based on the DVIS-DAQ framework [14]) achieves an AP of 49.7 and an $AR_{10}$ of 57.7, considerably exceeding the performance of the other methods. Switching from ResNet to a ViT backbone leads to a large performance gain, while our method still gives the best results, outperforming the underlying DVIS-DAQ model by 1.2 in terms of $AP_{50}$ and $AR_{10}$.

The OVIS dataset [5] is a more challenging dataset, with longer video sequences and occluded scenarios. In Table 3, we report the results on OVIS for the various methods (all, due to computational limitations, with ResNet-50 backbones). As we can see, we obtain the highest AP, $AP_{50}$ and $AR_1$ results.

As mentioned, one of the features of our MemCNP module is that it is model-agnostic and can thus be integrated into different VIS models. We consequently conduct experiments with different VIS frameworks enhanced by MemCNP and compare the obtained performance on YouTube-VIS to that of the underlying framework in Table 4. As can be seen from there, for the convolution-based framework SipMask, our method leads to AP improvements of 1.8 and 2.0 for ResNet-50 and ViT-b backbones, respectively. Similarly, MemCNP improves SeqFormer by 1.3 AP, and boosts IDOL by 3.2 (ResNet-50) and 2.8 (Swin-L), demonstrating its effectiveness with DETR-based models as well. Last but not least, MemCNP consistently enhances the performance of the MaskFormer DVIS-DAQ across all evaluation metrics.

Table 5 demonstrates the consistent improvements our proposed MemCNP achieves for the different VIS frameworks on the OVIS dataset. For SipMask, MemCNP yields significant improvements, boosting AP by 2.5 and $AP_{75}$ by 3.2. The improvements are more modest when applied to a stronger baseline like IDOL (+0.8 AP) or DVIS-DAQ (+0.3 AP) compared to the gains seen on YouTube-VIS (+2.8 and +1.2 AP). We attribute this to two factors. First, the primary challenge of OVIS is severe and prolonged occlusion, a problem that state-of-the-art models like DVIS-DAQ are already highly optimized to address. Second, our method's strength lies in leveraging uncertainty to identify reliable training samples. In a dataset dominated by heavily occluded—and thus inherently uncertain—instances, the pool of high-confidence samples that our method can exploit is naturally limited.

### 4.3. Ablation studies

We conduct a number of ablation studies on the YouTube-VIS dataset to demonstrate the importance of each component of our approach and to optimise its settings.

#### 4.3.1. Impact of individual components

We dissect the contribution of each key component of our method in Table 6. Our study follows a cumulative structure, starting with the SipMask baseline. First, we introduce our core probabilistic loss $L_{nll}$, which is fundamental to the MemCNP module and enables our Uncertainty Sample Selection (USS) strategy; yields an improvement of over 1.0 AP. Next, we incorporate the uncertainty regularisation term, $R_\sigma$, which further stabilizes training and boosts the AP by another 0.6 points. Finally, adding our Contrastive Instance Tracking (CIT) module yields an additional 0.4 AP gain, demonstrating the benefit of more discriminative instance features.

#### 4.3.2. Memory influence

We evaluate the performance of our method for different numbers of memory items. Intuitively, reducing the number of items results in a lack of prototypes, thereby downgrading the performance, while an excessively large number of items would lead to inefficient memory utilisation. The results presented in Table 7 confirm this, and show the median number of 128 memory items to give the best average precision of 40.1. To further assess the efficacy of our memory module, we apply a conventional CNP model [17] that aggregates context from fixed reference entries. As shown in Table 7, we find that the model with a learned memory module yields better performance for all memory settings, supporting the necessity of the memory module.

#### 4.3.3. Training rules

Looking at the performance using different training rules in MemCNP, reported in Table 8, the energy score introduced in [30] leads to an AP of 38.2. This is superseded by the vanilla NLL approach, which gives an AP of 39.7. The proposed Gaussian mixture model version of NLL leads to a further improvement, yielding the highest AP of 40.1 and $AR_{10}$ of 45.1. We conjecture that this discrepancy is due to the energy score enlarging the divergence between data points sampled from the generated distribution, which does not align with the objective of minimising detection variance, so that the two NLL-based losses significantly outperform it, while the GMM NLL performs slightly better than the vanilla NLL due to the local ensemble approach.

**Table 4**
VIS results on YouTube-VIS 2021 validation set, comparing the performance of different VIS frameworks with and without our proposed MemCNP.

| ResNet-based | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|
| SipMask | 27.5 | 47.8 | 27.9 | 26.4 | 31.9 |
| SipMask+MemCNP | **29.3** (+1.8) | **50.5** (+2.7) | **28.8** (+0.9) | **28.1** (+1.7) | **34.7** (+2.8) |
| SeqFormer | 37.7 | 58.3 | 41.6 | 34.8 | 46.2 |
| SeqFormer+MemCNP | **39.0** (+1.3) | **59.8** (+1.5) | **43.5** (+1.9) | **36.0** (+1.2) | **47.1** (+0.9) |
| IDOL | 43.9 | 68.0 | 49.6 | 38.0 | 50.9 |
| IDOL+MemCNP | **47.1** (+3.2) | **71.7** (+3.7) | **51.5** (+1.9) | **39.0** (+1.0) | **53.3** (+2.4) |
| DVIS-DAQ | 48.8 | 70.0 | 53.1 | 39.7 | 55.8 |
| DVIS-DAQ+MemCNP | **49.7** (+0.9) | **70.9** (+0.9) | **54.7** (+1.6) | **41.2** (+1.5) | **57.7** (+1.9) |
| **ViT-based** | **AP** | **$AP_{50}$** | **$AP_{75}$** | **$AR_1$** | **$AR_{10}$** |
| SipMask | 38.1 | 57.8 | 41.4 | 33.6 | 42.6 |
| SipMask+MemCNP | **40.1** (+2.0) | **59.7** (+1.9) | **44.5** (+3.1) | **35.9** (+2.3) | **45.1** (+2.5) |
| IDOL | 56.1 | 80.8 | 63.5 | 45.0 | 60.1 |
| IDOL+MemCNP | **58.9** (+2.8) | **82.4** (+1.6) | **65.0** (+1.5) | **46.8** (+1.8) | **63.6** (+3.5) |
| DVIS-DAQ | 60.4 | 81.2 | 68.1 | 46.9 | 66.1 |
| DVIS-DAQ+MemCNP | **61.6** (+1.2) | **83.0** (+1.8) | **69.6** (+1.5) | **48.4** (+1.5) | **67.3** (+1.2) |

**Table 5**
VIS results on OVIS validation set, comparing the performance of different VIS frameworks with and without our proposed MemCNP.

| ResNet-based | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|
| SipMask | 10.2 | 24.7 | 7.8 | 7.9 | 15.8 |
| SipMask+MemCNP | **12.7** (+2.5) | **27.5** (+2.8) | **11.0** (+3.2) | **8.1** (+0.2) | **16.5** (+0.7) |
| IDOL | 29.5 | 50.1 | 29.4 | 14.9 | 37.1 |
| IDOL +MemCNP | **30.3** (+0.8) | **55.6** (+5.5) | **30.2** (+0.8) | **14.7** (-0.2) | **39.4** (+2.3) |
| DVIS-DAQ | 37.4 | 63.2 | 37.8 | 16.7 | 44.5 |
| DVIS-DAQ+MemCNP | **37.7** (+0.3) | **64.3** (+1.1) | **37.7** (-0.1) | **16.9** (+0.2) | **44.3** (-0.2) |

**Table 6**
Ablation results, on YouTube-VIS 2021, for different components of our approach.

| $\mathcal{L}_{nll}$ | $R_\sigma$ | CIT | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|---|---|
| – | – | – | 38.1 | 57.8 | 41.4 | 33.6 | 42.6 |
| ✓ | – | – | 39.1 | 57.4 | 42.4 | 34.9 | 44.5 |
| ✓ | ✓ | – | 39.7 | **62.1** | 43.3 | 34.4 | 44.4 |
| ✓ | ✓ | ✓ | **40.1** | 59.7 | **44.5** | **35.9** | **45.1** |

**Table 7**
YouTube-VIS 2021 results for different numbers of memory items in MemCNP.

| Memory items | Type | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|---|
| 64 | CNP | 36.7 | 57.2 | 40.8 | 33.3 | 41.8 |
| 128 | CNP | 38.7 | 58.6 | 41.5 | 34.6 | 42.7 |
| 256 | CNP | 34.6 | 55.1 | 37.2 | 31.6 | 38.6 |
| 64 | MemCNP | 38.9 | 57.7 | 43.1 | 34.9 | 44.3 |
| 128 | MemCNP | **40.1** | **59.7** | **44.5** | **35.8** | **45.1** |
| 256 | MemCNP | 38.7 | 58.7 | 41.5 | 34.7 | 42.8 |

**Table 8**
YouTube-VIS 2021 results for different training rules in MemCNP.

| Training rule | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|
| - | 38.1 | 57.8 | 41.4 | 33.6 | 42.6 |
| energy score [30] | 38.2 | 56.8 | 42.0 | 34.0 | 42.7 |
| vanilla NLL | 39.7 | **61.4** | 42.6 | 35.2 | 44.7 |
| GMM NLL | **40.1** | 59.7 | **44.5** | **35.9** | **45.1** |

#### 4.3.4. Sample selection strategy

We evaluate different sample selection strategies, and report the results in Table 9. The strategies in YOLO [43] and FCOS [16] tend to favour larger objects over smaller ones due to the higher likelihood of obtaining more training candidates based on their criteria (see Fig. 4). This results in relatively lower APs of 38.4 and 38.9, respectively, since the YouTube-VIS dataset comprises objects of various spatial sizes. As we can see, our proposed USS strategy outperforms not only YOLO's and FCOS's approaches but also that used in ObjectBox [42], by 0.6 in AP and 0.9 in $AR_1$, respectively. The reasons for this are two-fold. First, we incorporate a stricter criterion for selecting not only positive candidates but also trustworthy negatives. Second, prioritising centredness scores can result in the selection of false positives, where objects that are not located at the centre are mistakenly chosen, as happens in feature stage 3 of the ObjectBox example in Fig. 4. In contrast, our USS can dynamically handle this by considering uncertainties.

**Table 9**
YouTube-VIS 2021 results for different sample selection strategies.

| Strategy | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|
| YOLO | 38.4 | 59.5 | 41.9 | 34.6 | 43.5 |
| FCOS | 38.9 | 59.8 | 42.0 | 33.9 | 44.2 |
| ObjectBox | 39.5 | **60.3** | 43.6 | 35.0 | 44.5 |
| USS | **40.1** | 59.7 | **44.5** | **35.9** | **45.1** |

### 4.4. Qualitative results

#### 4.4.1. Uncertainty visualisation

In Fig. 6, we present two examples to illustrate the predictive capability of the uncertainty maps and to demonstrate why our method relies on them to select reliable training samples. One of the frames contains two objects that have minimal overlap, while the other example exhibits significant overlap.

As we can see from Fig. 6, for the case of slight overlapping, lower uncertainties (or higher certainties) tend to accumulate closer to the central locations of each object across all stages. In contrast, when objects overlap significantly, lower certainties are more likely to appear

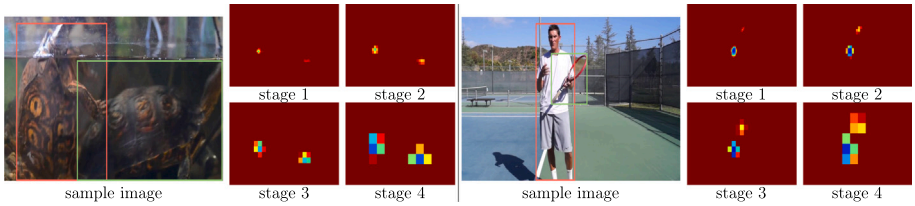


**Fig. 6.** Two sample frames from YouTube-VIS 2021 with their associated uncertainty maps. The maps from the different features stages are scaled to the same size for clarity; warmer colours indicate higher uncertainty, cooler colours higher certainty..

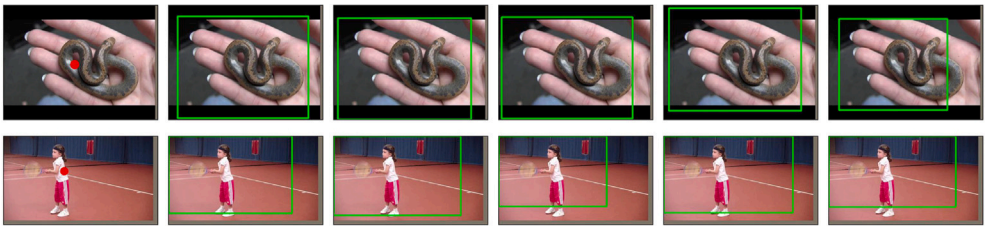

**Fig. 7.** Sample visualisations of the top five most relevant memories for a given query position from YouTube-VIS 2021. From left to right: query, top-1, top-2, top-3, top-4, and top-5.

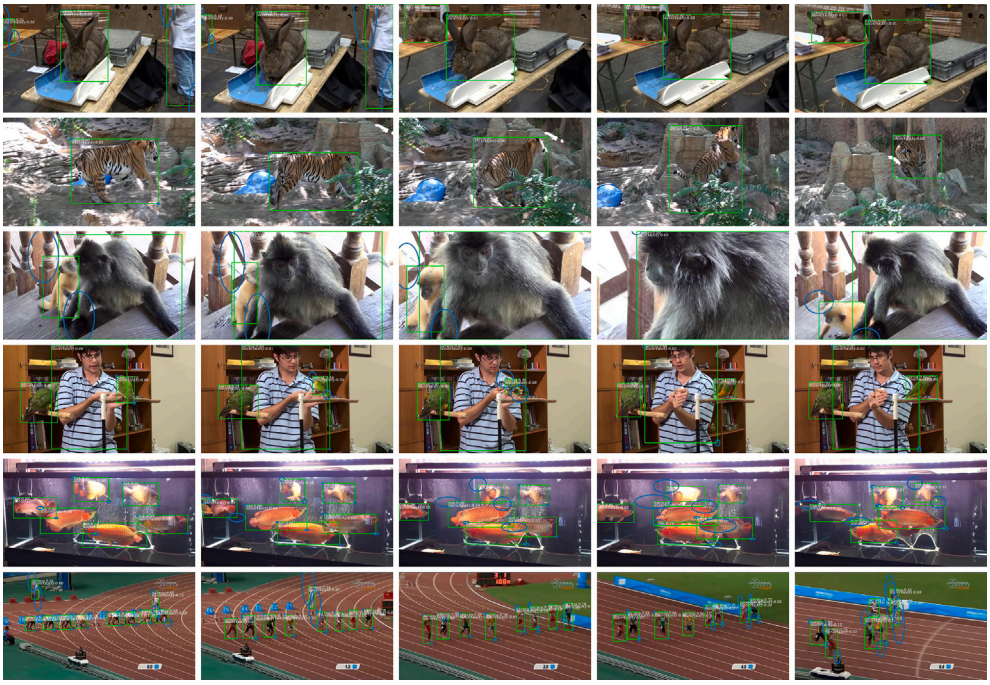

**Fig. 8.** Sample visualisations of detection results on YouTube-VIS 2021 validation set. The size of the ellipses corresponds to the uncertainty magnitudes. Zoom in for best view.

in regions where the objects genuinely exist rather than in overlapping areas. Consequently, by leveraging the information provided by the uncertainty maps, our method is capable of dynamically identifying the most trustworthy sample locations for subsequent training objectives, even in complex scenarios.

#### 4.4.2. Memory visualisation

We investigate what resides in the memory bank by passing the image into the encoder, obtaining the latent representation $Z$. Then a latent vector is randomly sampled to retrieve the top 5 most relevant memory items through Eq. (5). Subsequently, we translate the tail dimensions of the retrieved memories into four distances, plotted as bounding boxes. These boxes illustrate the probable distribution of the object relative to the query location, as depicted in Fig. 7. We can observe that even without involving the final regression head, the retrieved memory demonstrates to some extent the ability to localise the object corresponding to the query representation.

#### 4.4.3. Probabilistic detection results

We visualise some of our detection results in Fig. 8, together with the corresponding uncertainties for each object (in form of the size of the ellipses). In the top two examples, where the objects are well-represented with minimal occlusion and blurring, our method successfully assigns small uncertainties to accurately recognised objects, except

for partially occluded ones. In rows 3 and 4, objects with high occlusion are effectively detected, with larger ellipses representing their corresponding uncertainties, while in the bottom two examples, which contain more object instances, our method assigns higher uncertainties to those with irregular appearances (row 5) and those that are farther away (row 6).

#### 4.4.4. Video instance segmentation results

We demonstrate some final video instance segmentation results in Fig. 9, where different colours represent different instances. Examining the top four examples, where instances have minimal overlaps, we can observe that the target objects are accurately detected, segmented, and tracked across video frames. In more challenging scenarios, where multiple object instances frequently change appearance and position, our method still achieves satisfactory results. Notably, in rows 5 and 6, objects are clearly segmented without confusion. Moreover, as can be seen from the bottom two examples, our method continues to function effectively even in presence of high uncertainties in the detected objects.

#### 4.4.5. Failure cases and limitations

In Fig. 10, we illustrate two failure cases to analyse the limitations of our method. In the example on the left, the baseline method fails to detect the bottom right giraffe while our MemCNP approach consistently detects and segments it. However, neither method precisely identifies the giraffe due to its similar appearance to other instances in this highly overlapping scenario. The example on the right presents a similar issue where two monkeys overlap when interchanging their positions. Highly similar or overlapping objects can lead to less discriminative embeddings, making instance differentiation challenging. Also,

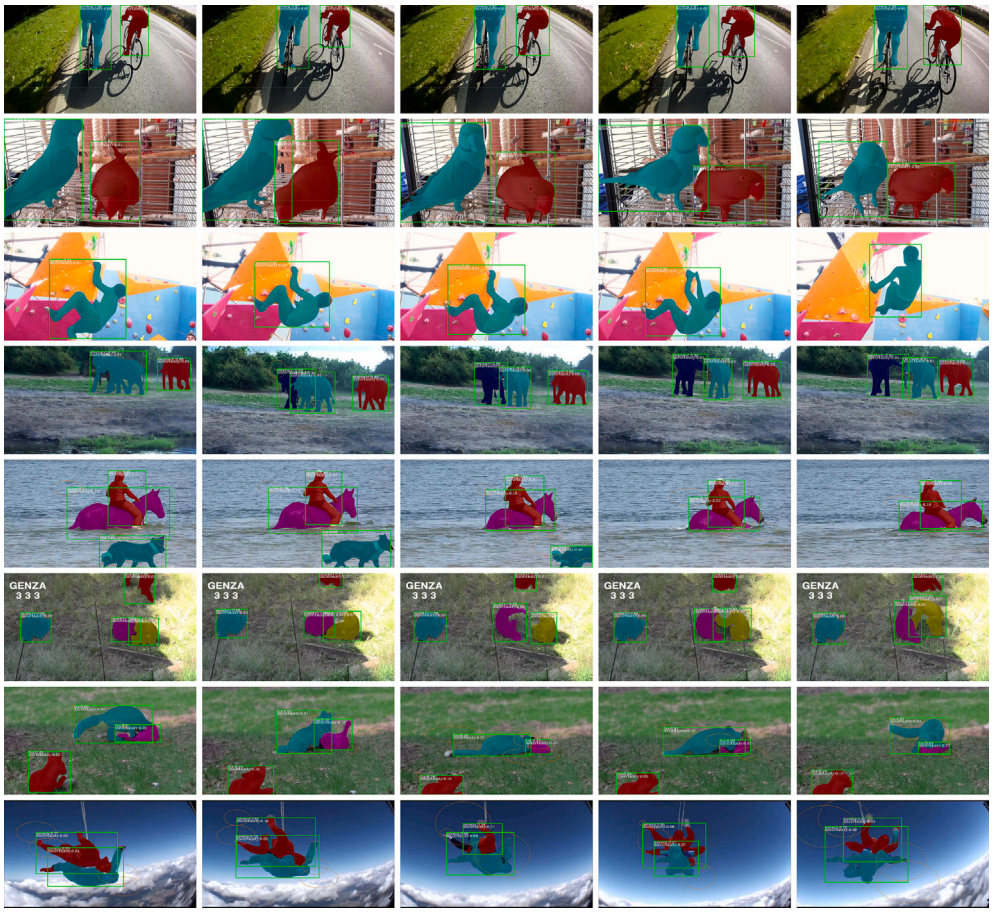


**Fig. 9.** Examples of video instance segmentation results on YouTube-VIS 2021 validation set. Zoom in for best view.

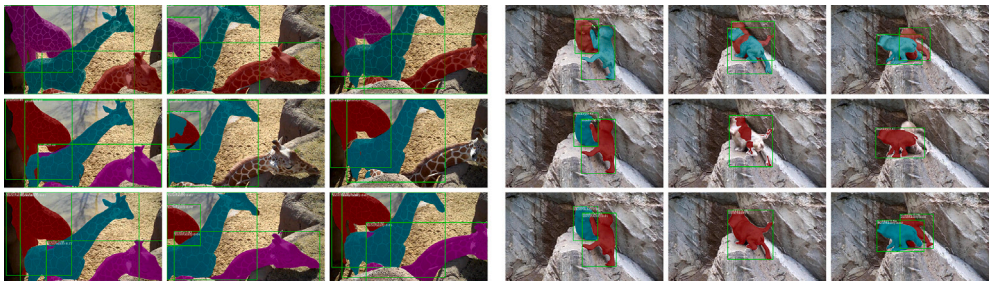

**Fig. 10.** Exemplar failure cases from YouTube-VIS 2021. From top to bottom, the images are generated from ground truth, baseline, and MemCNP, respectively. Zoom in for best view.

when similar objects swap positions or reappear after occlusion, tracking inconsistencies arise. While MemCNP improves reliability, it does not explicitly model long-term object associations, leading to potential error accumulation. Stronger contrastive learning or transformer-based attention across longer time horizons [21] could potentially help in these cases.

## 5. Conclusions

In this paper, we have proposed a novel approach to video instance segmentation that integrates memory networks into conditional neural processes for better VIS performance. By leveraging the benefits of probabilistic modelling, we obtain uncertainties associated with network predictions to improve the reliability of VIS models. Furthermore, this enables us to dynamically select more reliable samples for multi-task training in the VIS task, while we also incorporate a contrastive learning strategy to enhance instance tracking by learning more distinctive features of individual instances. Extensive experiments on the YouTube-VIS and OVIS datasets convincingly demonstrate the effectiveness of our proposed method, showing not only significant improvements on the underlying baseline models, but also gaining valuable insights into the reliability of model predictions. Notably, our MemCNP is designed to be model-agnostic and can thus be applied in various VIS frameworks.

In future work, apart from enhancing failure prediction, we plan to further investigate the impact of varying video resolutions and frame rates to assess the robustness of our method and to evaluate the generalisation and reliability of our approach on other datasets such as the Video Panoptic Segmentation (VPS) dataset [48] as well as data from other domains. Moreover, we plan to extend MemCNP with spatio-temporal transformers [21] or adopt temporal-level memories for longer video clips to leverage more temporal information while maintaining model efficiency.

## CRediT authorship contribution statement

**Kunhao Yuan:** Writing – review & editing, Writing – original draft, Visualisation, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gerald Schaefer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Yu-Kun Lai:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **Xiyao Liu:** Writing – review & editing, Validation, Formal analysis. **Lin Guan:** Validation, Supervision, Formal analysis. **Hui Fang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This research is supported by Natural Science Foundation of Hunan Province, China (2022GK5002, 2024JK2015, 2024JJ5440), Special Foundation for Distinguished Young Scientists of Changsha (kq2209003).

## Data availability

I have shared the link to the data and code used in this article.

## References

[1] L. Yang, Y. Fan, N. Xu, Video instance segmentation, in: IEEE International Conference on Computer Vision, 2019, pp. 5188–5197.
[2] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2961–2969.
[3] J. Cao, R.M. Anwer, H. Cholakkal, F.S. Khan, Y. Pang, L. Shao, SipMask: spatial information preservation for fast image and video instance segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 1–18.
[4] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, X. Bai, In defense of online models for video instance segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 588–605.
[5] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P.H.S. Torr, S. Bai, Occluded video instance segmentation: a benchmark, Int. J. Comput. Vis. 130 (8) (2022) 2022–2039.
[6] J. Wu, Y. Jiang, S. Bai, W. Zhang, X. Bai, SeqFormer: sequential transformer for video instance segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 553–569.
[7] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, B. Leibe, STEM-Seg: spatio-temporal embeddings for instance segmentation in videos, in: European Conference on Computer Vision, Springer, 2020, pp. 158–177.
[8] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, H. Xia, End-to-end video instance segmentation with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 8741–8750.
[9] M. Heo, S. Hwang, S.W. Oh, J.-Y. Lee, S.J. Kim, Vita: video instance segmentation via object token association, in: Neural Information Processing Systems, 2022.
[10] J. Wu, S. Yarram, H. Liang, T. Lan, J. Yuan, J. Eledath, G. Medioni, Efficient video instance segmentation via tracklet query and proposal, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 959–968.
[11] M. Heo, S. Hwang, J. Hyun, H. Kim, S.W. Oh, J.-Y. Lee, S.J. Kim, A generalized framework for video instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 14623–14632.
[12] A. Choudhuri, G. Chowdhary, A.G. Schwing, Context-aware relative object queries to unify video instance and panoptic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 6377–6386.
[13] Y. Tian, H. Fu, H. Wang, Y. Liu, Z. Xu, H. Chen, J. Li, R. Wang, Rgb oralscan video-based orthodontic treatment monitoring, Sci. China Inf. Sci. 67 (1) (2024) 112107.
[14] Y. Zhou, T. Zhang, S. Ji, S. Yan, X. Li, Improving video segmentation via dynamic anchor queries, in: European Conference on Computer Vision, Springer, 2024, pp. 446–463.
[15] Y. Tian, G. Cheng, J. Gelernter, S. Yu, C. Song, B. Yang, Joint temporal context exploitation and active learning for video segmentation, Pattern Recognit. 100 (2020) 107158.
[16] Z. Tian, C. Shen, H. Chen, T. He, FCOS: a simple and strong anchor-free object detector, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2020) 1922–1933.
[17] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y.W. Teh, D. Rezende, S.M.A. Eslami, Conditional neural processes, in: International Conference on Machine Learning, PMLR, 2018, pp. 1704–1713.
[18] L. Yang, Y. Fan, Y. Fu, N. Xu, The 3rd large-scale video object segmentation challenge - video instance segmentation track (Jun. 2021).
[19] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. Van Gool, The 2017 davis challenge on video object segmentation, arXiv preprint arXiv:1704.00675, 2017.
[20] D. Bolya, C. Zhou, F. Xiao, Y.J. Lee, YOLACT: real-time instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9157–9166.
[21] Y. Tian, Y. Zhang, D. Zhou, G. Cheng, W.-G. Chen, R. Wang, Triple attention network for video segmentation, Neurocomputing 417 (2020) 202–211.
[22] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L.Y. Wu, Y. Liu, C. Fan, Y. Zhuge, C. Shen, Ctvis: consistent training for online video instance segmentation, in: IEEE International Conference on Computer Vision, 2023, pp. 899–908.
[23] H. Fang, T. Zhang, X. Zhou, X. Zhang, Learning better video query with sam for video instance segmentation, IEEE Trans. Circuits Syst. Video Technol. 35 (4) (2024) 2963–2974.
[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[25] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, A.G. Schwing, Mask2former for video instance segmentation, arXiv preprint arXiv:2112.10764, 2021.
[26] D.J.C. MacKay, Bayesian neural networks and density networks, Nucl. Instrum. Methods Phys. Res. Sect. A 354 (1) (1995) 73–80.
[27] Y. Jeon, G. Hwang, Bayesian mixture of gaussian processes for data association problem, Pattern Recognit. 127 (2022) 108592.
[28] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: International Conference on Learning Representations, 2013, https://api.semanticscholar.org/CorpusID:216078090.
[29] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Neural Inf. Process. Syst. 30 (2017).
[30] A. Harakeh, S.L. Waslander, Estimating and evaluating regression predictive uncertainty in deep object detectors, in: International Conference on Learning Representations, 2021.
[31] G. Hess, C. Petersson, L. Svensson, Object detection as probabilistic set prediction, in: European Conference on Computer Vision, Springer, 2022, pp. 550–566.
[32] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D.J. Rezende, S.M. Eslami, Y.W. Teh, Neural processes, arXiv preprint arXiv:1807.01622, 2018.
[33] H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, Y.W. Teh, Attentive neural processes, in: International Conference on Learning Representations, 2019.
[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[35] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine

translation, in: Conference on Empirical Methods in Natural Language Processing, 2014.
[36] J. Weston, S. Chopra, A. Bordes, Memory networks, in: International Conference on Learning Representations, 2014.
[37] C. Chunseong Park, B. Kim, G. Kim, Attend to you: personalized image captioning with context sequence memory networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 895–903.
[38] H. Huang, M. Luo, R. He, Memory uncertainty learning for real-world single image deraining, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 3446–3460.
[39] J. Miao, Y. Wei, Y. Yang, Memory aggregation networks for efficient interactive video object segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10366–10375.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al, An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 10012–10022.
[42] M. Zand, A. Etemad, M. Greenspan, Objectbox: from centers to boxes for anchor-free object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 390–406.
[43] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, arXiv preprint arXiv:1804.02767, 2018.
[44] X. Li, J. Wang, X. Li, Y. Lu, Video instance segmentation by instance flow assembly, IEEE Trans. Multimed. 25 (2022) 7469–7479.
[45] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[46] K. Yuan, G. Schaefer, Y.-K. Lai, Y. Wang, X. Liu, L. Guan, H. Fang, A multi-strategy contrastive learning framework for weakly supervised semantic segmentation, Pattern Recognit. 137 (2023) 109298.
[47] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[48] J. Miao, X. Wang, Y. Wu, W. Li, X. Zhang, Y. Wei, Y. Yang, Large-scale video panoptic segmentation in the wild: a benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 21033–21043.

## Author biography

**Kunhao Yuan** received his B.Sc. from Northeastern University, China in 2017 and his M.Sc. and Ph.D. from Loughborough University in 2018 and 2023, respectively. From later 2018 to early 2020, he was an algorithmic engineer with UnionBigData, China. He is now a postdoctoral research fellow at the University of Edinburgh. His research interests include computer vision, machine intelligence and computational neuroscience.

**Gerald Schaefer** gained his PhD in Computer Vision from the University of East Anglia. He worked at the Colour & Imaging Institute, University of Derby, in the School of Information Systems, University of East Anglia, in the School of Computing and Informatics at Nottingham Trent University, and in the School of Engineering and Applied Science at Aston University before joining the Department of Computer Science at Loughborough University. His research interests are mainly in the areas of computer vision, colour image analysis, medical imaging, and computational intelligence. He has published extensively in these areas with a total publication count of about 500, has been invited as keynote or tutorial speaker to numerous conferences, is the organiser of various international workshops and special sessions at conferences, and the editor of several books, conference proceedings and special journal issues.

**Yu-Kun Lai** is a Professor at School of Computer Science and Informatics, Cardiff University, UK. He received his bachelor's and PhD degrees in Computer Science from Tsinghua University, in 2003 and 2008 respectively. His research interests include computer graphics, computer vision, geometric modeling and image processing. He is on the editorial boards of Computer Graphics Forum and The Visual Computer. For more information, visit https://users.cs.cf.ac.uk/Yukun.Lai

**Xiyao Liu** was born in Hunan Province, in 1987. He received the B.S. degree and the Ph.D. degree from School of Electrical Engineering and Computer Sciences, Department of Microelectronics, Peking University in 2008 and 2015. He is now an associate professor in School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include information security, multimedia technology, computational vision, and deep learning.

**Lin Guan** is a Senior Lecturer in Computer Science at Loughborough University. Her research focuses on performance modelling of heterogeneous networks and systems, QoX-QoS analysis, edge/fog computing, VANETs, SDN, cloud/mobile computing, wireless sensor networks, multimedia systems, and Model-Based Systems Engineering (MBSE) with QoS attributes. She has published over 100 papers and serves on editorial boards for top journals, including Elsevier's Journal of Systems and Software and Simulation Modelling Practice and Theory. She has received several prestigious awards, including a Royal Society Industry Fellowship and EPSRC-funded projects, and currently leads an EPSRC/Rolls-Royce MBSE project.

**Hui Fang** received the B.S. degree from the University of Science and Technology, Beijing, China, in 2000 and the Ph.D. degree from the University of Bradford, U.K., in 2006. He is currently with the Computer Science Department at Loughborough University. Before, he has carried out research at several world-leading universities, such as University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualisation, visual analytics, and artificial intelligence. During his career, he has published more than 100 journal and conference papers.