MINet: Meta-Learning Instance Identifiers for Video Object Detection

Jiajun Deng, *Graduate Student Member, IEEE*, Yingwei Pan, Ting Yao, *Member, IEEE*, Wengang Zhou, *Senior Member, IEEE*, Houqiang Li, *Fellow, IEEE*, and Tao Mei, *Fellow, IEEE*

***Abstract*— Recent advances in video object detection have characterized the exploration of temporal coherence across frames to enhance object detector. Nevertheless, previous solutions either rely on additional inputs (e.g., optical flow) to guide feature aggregation, or complex post-processing to associate bounding boxes. In this paper, we introduce a simple but effective design that learns instance identifiers for instance association in a meta-learning paradigm, which requires no auxiliary inputs or post-processing. Specifically, we present Meta-Learnt Instance Identifier Networks (namely MINet) that novelly meta-learns instance identifiers to recognize identical instances across frames in a single forward-pass, leading to the robust online linking of instances. Technically, depending on the detection results of previous frames, we teach MINet to learn the weights of an instance identifier on the fly, which can be well applied to up-coming frames. Such meta-learning paradigm enables instance identifiers to be flexibly adapted to novel frames at inference. Furthermore, MINet writes/updates the detection results of previous instances into memory and reads from memory when performing inference to encourage temporal consistency for video object detection. Our MINet is appealing in the sense that it is pluggable to any object detection model. Extensive experiments on ImageNet VID dataset demonstrate the superiority of MINet. More remarkably, by integrating MINet into Faster R-CNN, we achieve 80.2% mAP on ImageNet VID dataset.**

***Index Terms*— Video object detection, meta learning, memory network, box association.**

I. Introduction

OBJECT detection is one of fundamental problems in computer vision field, which aims to identify objects within images and spatially localize them with bounding boxes. The recent development of deep learning [1]–[5] has successfully pushed the limits of object detection, leading to a surge of deep object detectors [6]–[14] that follow the typical region-based detection paradigm. In a further step to recognize and localize objects in videos, video object detection [15]–[18] has extended the visual perception from individual image/frame to a sequence of frames. Nevertheless, considering that video is an information-intensive media with both spatial and temporal complexities, simply applying image object detectors will suffer from unsatisfactory robustness especially when target objects undergo motion blur or occlusions. Such facts motivate the exploration of spatio-temporal coherence in videos to boost object detectors.

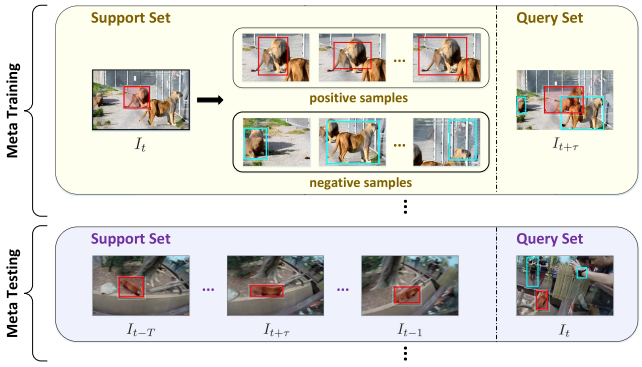


Fig. 1. Meta-learning for instance association across frames in video object detection. The red boxes indicate the target object and positive samples, and the blue boxes represent non-associated ones. At meta-training stage, we randomly sample regions from different frames to construct support and query sets. During meta-testing, we take samples from history frames as the support set and take the region proposals from current frame as the query set.

In the literature, there have been a series of innovations being proposed to exploit such temporal coherence across frames for video object detection. Two representative research directions are feature aggregation [18]–[20] and box-level association [15]–[17]. The former leverages additional inputs (*e.g.*, optical flow) to guide spatio-temporal aggregation of nearby frames and thus enhances per-frame features for detection. Such design is sometimes unsteady if the auxiliary optical flow inputs are deteriorated by motion blur or occlusion. The latter commonly associates bounding boxes (*i.e.*, instances) across consecutive frames through post-processing via linking or tracking. However, the independent process of box-level association may destroy the interaction between localizing and associating instances within videos, resulting in a sub-optimal solution. In contrast, we propose to mitigate these issues by learning instance identifier for instance association in a meta-learning paradigm. We devise an unified framework for both intra-frame object detection and inter-frame object association, and this architecture works without any auxiliary input or post-processing. Figure 1 depicts the meta-learning

Manuscript received June 15, 2020; revised April 15, 2021 and July 8, 2021; accepted July 13, 2021. Date of publication July 30, 2021; date of current version August 5, 2021. This work was supported in part by the National Key Research and Development Program of China under contract 2017YFB1002202, in part by the National Natural Science Foundation of China under Contract 61836011 and Contract 62021001, in part by the Youth Innovation Promotion Association Chinese Academy of Sciences (CAS) under Grant 2018497, and in part by the Graphics Processing Unit (GPU) cluster built by Multimedia Computation and Communication Laboratory (MCC) Lab of Information Science and Technology Institution, USTC. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mireille Boutin. *(Corresponding authors: Yingwei Pan; Wengang Zhou.)*

Jiajun Deng, Wengang Zhou, and Houqiang Li are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei 230026, China (e-mail: dengjj@mail.ustc.edu.cn; zhwg@ustc.edu.cn; lihq@ustc.edu.cn).

Yingwei Pan, Ting Yao, and Tao Mei are with JD AI Research, Beijing 100105, China (e-mail: panyw.ustc@gmail.com; tingyao.ustc@gmail.com; tmei@jd.com).

Digital Object Identifier 10.1109/TIP.2021.3099409

setup for cross-frame instance association in this work. The standard meta-training stage involves iterating through training episodes [21]–[23], each of which consists of a support set and a query set. We formulate each training episode as a new task of instance association. Specifically, as shown in the top row of this figure, to generate samples of the support set, we perform random perturbation over the ground-truth box of a chosen object in frame I_t. The samples of the query set are generated in the same way in frame $I_{t+\tau}$. The objective of meta-training is to leverage samples in support set to learn instance identifiers that can be applied to the query set of up-coming frame with good performance. Similarly, we follow the meta-testing process to perform box association at the inference stage. Particularly, during inference, the samples of the support set are generated with the history detections, and the region proposals of the up-coming frame are regarded as the samples of the query set. Consequently, such meta-learning paradigm leads to more robust online association of instances.

By consolidating the idea of meta-learning instance identifiers for instance association and empowering the joint training of object detection and association, we present a new Meta-Learnt Instance Identifier Networks (MINet) to enhance video object detection. Specifically, a standard image object detector (*i.e.*, Faster R-CNN) is adopted to localize objects within each frame. In the meantime, we take ground-truth boxes in nearby frames as support set in meta-training stage, which will be exploited to predict the weights of an instance identifier on the fly. The obtained instance identifier is applied to identify the same instance within the up-coming query frame. In this sense, the instance association is naturally built between nearby frames and query frame. The whole architecture is end-to-end optimized by minimizing the detection loss over each frame and the error of identifying instance labels within query frame conditioned on the support frames. At inference stage, we sequentially construct the support set with detected boxes in previous frames. An additional dynamic memory is utilized to manage the history of the detection results for instances in sequential support set, from which MINet reads to enhance the detection results of up-coming frame via instance association in between. With these designs, our proposed MINet can make temporal consistent predictions of consecutive frames in a single forward-pass. After that, the dynamic memory is updated according to the enhanced detection results.

In summary, we make three-fold contributions:

- We propose a novel architecture to address the issue of video object detection, which meta-learns instance identifiers for instance association to make consistent and robust detection predictions.
- We present an elegant view of how to learn the parameters of instance identifiers on the fly, and how to enhance temporal coherence across frames at inference, which are problems not yet fully studied.
- We conduct extensive experiments to validate the merits of our method, and show encouraging improvement over previous methods following the box-level association paradigm.

II. Related Work

A. Object Detection in Images

The development of deep convolutional neural networks [1]–[5], [24] has witnessed recent advances in object detection technologies [6]–[11], [25]–[31]. Generally, research in object detection has proceeded along two directions: two stage detectors based on region proposals and one stage detectors that directly applied on feature maps.

Two stage paradigm is first introduced in R-CNN [9] by making detections on deep features of region proposals generated from Selective Search [32]. SPPnet [10] and Fast R-CNN [8] propose spatial pyramid pooling or ROI pooling to get rid of redundancy in region feature extraction. Later on, Faster R-CNN [13] devises Region Proposal Network (RPN) to replace Selective Search for region proposal generation with crafted anchor boxes of different scales and aspect ratios in each feature grid. More recently, further improvements are achieved by variants of R-CNN series [6], [12], [33]–[35]. In addition, inspired by domain adaptation [36], [37] for recognition, [38], [39] focus on learning robust and domain-invariant detectors based on two stage paradigm.

In the one stage direction, bounding boxes are directly predicted by sliding windows without the stage of region proposal. The early successes of one stage object detectors (*e.g.*, SSD [26], YOLOv2/v3 [28], [40], and RetinaNet [25]) mainly utilize pre-defined anchor boxes on multi-scale feature maps to enumerate possible location of objects. Recently, anchor-free object detectors [41]–[44] are revised to improve the generalization ability and avoid extra hyper-parameters (*i.e.*, the design of anchor boxes and the assignment of positive anchors). In this work, we adopt the widely-adopted Faster R-CNN as the base object detector for each single frame.

B. Object Detection in Videos

The still image object detectors operating on single frames suffer from drastic confidence fluctuations in video applications, which motivates the research to incorporate temporal context for object detection in videos. Video object detection algorithms can be divided into two categories: feature-level aggregation based methods [18], [19], [45]–[51] and box-association based methods [15]–[17], [52], [53]. The methods of the first category generally improve per-frame features by taking the aggregation of adjacent frames. Within this type of methods, FGFA [18] utilizes optical flow as guidance to take motion compensation on feature maps of nearby frames and aggregates them into the target one. MANet [20] re-utilizes the optical flow to further predict the movements of boxes, and extends the pixel-level feature calibration proposed in FGFA with box-level detection score calibration. STSN [45] exploits deformable convolution layers [7] for spatiotemporal sampling across frames with self-learnt sampling offset. PSLA [46] proposes a novel Progressive Sparse Local Attention to establish sparse correspondence between pixels across frames in a local region to get rid of optical flow prediction. OGEM [47] takes the hard-attention of objects as guidance to construct a storage-efficient memory block for feature aggregation. Compared to previous works that

aggregate features in pixel-level, RDN [50] extends object-to-object relation reasoning into video domain and augments the per-region features by a set of supportive proposals from adjacent frames. Similar to RDN, SELSA [54] also performs feature aggregation in proposal-level, but aggregate semantic features from the full-sequence instead of adjacent frames. MEGA [55] introduces memory enhanced global-local aggregation mechanism to take full consideration of both global and local information.

In the other direction, box-level association methods associate per-frame detection boxes into tubelets and re-score them in each tube for refinement. In the early stage, T-CNN [53] proposes motion-guided propagation to recover false negatives by detections from adjacent frames and high confidence tracking to generate long term tubelet for re-scoring. Seq-NMS [16] boosts scores of weaker predictions by high-scoring ones from adjacent frames before applying NMS. D&T [15] jointly performs object detection and tracking, which links tracklets to object tubes and find the optimal tube for score promotion in post-processing. TCD [56] proposes to tightly integrate object detection and multiple object tracking by conditioning object detection on the previous tracklets. ST-lattice [52] devises a sophisticated system to adaptively perform expensive detection and cheaper propagation across scale/time with multiple networks.

C. Meta-Learning and Its Application in Videos

Meta-learning aims to train a model on a variety of learning tasks, and enable the model to deal with new learning tasks with a small number of samples. In object detection and video application, there are a few works with meta-learning paradigm. MetaDet [23] leverages the meta-level knowledge from a large quantity of data with base classes to facilitate model generation to detect novel classes. Meta-Tracker [57] learns to improve the robustness of initial target model by leveraging the error signals from the future frames through meta-training. Retina-MAML and FCOS-MAML proposed in [58] follow a three-step schedule with offline model-agnostic meta-training (MAML [21]) to convert general object detectors (*i.e.*, RetinaNet [25] and FCOS [42]) into trackers. Unlike these techniques that were developed for image object detection and visual object tracking, the meta-learning paradigm in our MINet is applied for video object detection.

D. Memory Networks and Its Applications in Videos

Temporal coherence is widely exploited for addressing the issue of video applications [59]–[67]. COSNet [59] introduces a global co-attention mechanism and an alternated network training strategy to facilitate segmenting foreground objects in unsupervised settings. AGNN [60] devises novel video graphs to re-formulate zero-shot video object segmentation as a process of iterative information fusion. Within the approaches of leveraging temporal coherence, there have been several innovations being proposed to exploit memory networks in video applications. MemTrack [61] devises a dynamic memory network to adapt matching templates to variable object appearance for visual tracking. STM [62] develops a memory network to record the cues of past frames as an external memory, and learns to read the recorded information to ameliorate object segmentation in the current frame. GraphMemVOS [63] introduces a novel graph memory mechanism that generates video-specific memory summarization in an episodic manner to adapt the network for benefiting video object segmentation. MA-Net [64] leverages a novel memory aggregation mechanism to improve the performance of interactive video object segmentation by aggregating the information of previous interaction rounds. LFB [68] records the supportive information over the entire span of a video with a long-term feature bank to facilitate action detection. In this work, we also devise memory networks to facilitate model inference. However, different from the previous methods that perform model update [61], [63] or feature enhancement [62], [64], [68] with spatio-temporal features from the memory, we record the history results in each memory slot, and perform online box association to improve the robustness of single-frame detections.

E. Summary

Our MINet belongs to box-level association methods. In contrast to existing techniques that perform object linking algorithms as post processing [15], [16], [53] or with multiple networks [52], we novelly introduce a unified architecture for object detection and cross-frame instance association via the designed meta-learnt instance identifiers. Moreover, at inference, we capitalize on a dynamic memory module to perform object detection and box-to-memory association in a single forward-pass. Particularly, our proposed MINet share the similar spirits with RDN to improve the robustness of video object detection in box-level when performing online inference. However, RDN exploits the attention mechanism to augment the target proposals with supportive proposal features from adjacent frames, while our MINet devises meta-learnt instance identifiers to associate per-frame detections out of the base detector. These two methods target to tackle the same problem, but following different paradigms.

III. MINet for Video Object Detection

We devise Meta-Learning Instance Identifiers Networks (MINet) to facilitate robust object detection in videos by casting object detection and association within a unified architecture. Specifically, MINet meta-learns instance identifiers during meta-training stage, whose parameters are produced on the fly conditioned on existing detection results of previous frames. The meta-learnt instance identifiers are applied for up-coming query frames, enabling online association of instances across frames. Such meta-training stage can be flexibly integrated into the training process of general object detection models by attaching a sibling branch.

During inference, we capitalize on a dynamic memory module to trace the footprint of existing detected objects. Concretely, the detection results of previous frames are written as memorized objects, which will be further leveraged to rectify the predictions of detected instances in the up-coming frame via box-to-memory association. Such a process is sequentially conducted after updating memorized items with

TABLE I
THE MAIN ACRONYMS AND NOTATIONS USED IN THIS PAPER

t, $t-1$, $t+\tau$	video frame indices
L	the total length of a video sequence
$\dot{I}_t$	the frame to be detected at present
$\{I_t\}$	a set of consecutive video frames
$\boldsymbol{\zeta}_Q$, $\boldsymbol{\zeta}_S$,	the query/support set for meta-learning
N_Q, N_S	the number of samples in the query/support set
A_Q, A_S	the association embedding matrix consists of samples in the query/support set
InsID	the instance identification label, '1' for associated instances and '0' for others
$\mathcal{P}_t$	the set of base detections for frame I_t
N_P	the number of base detections
p_t^n	the n-th detected object from $\mathcal{P}_t$
$\boldsymbol{b}_t^n$	the box coordinates of p_t^n
$\boldsymbol{s}_t^n$	the classification score of p_t^n
$\boldsymbol{a}_t^n$	the association embedding of p_t^n
$\mathcal{M}_{t-1}$	the set of recorded memory slots updated up to frame I_{t-1}
K	the number of valid memory slots
m_{t-1}^k	the k-th memory slot from $\mathcal{M}_{t-1}$
$\boldsymbol{\pi}_{t-1}^m$	the memory score of m_{t-1}^k
$\boldsymbol{\gamma}_{t-1}^k$	the meta-learnt instance identifier of m_{t-1}^k
$\boldsymbol{\zeta}_{t-1}^k$	a memory buffer with fixed capacity to store the support samples for updating $\boldsymbol{\gamma}_{t-1}^k$

rectified predictions. The main acronyms and notations used in this paper are summarized in Table I for better understanding.

A. Overview

Given a snippet of video with L consecutive frames $\{I_t\}_{t=1}^L$, the task of video object detection is to localize and recognize objects in each frame by additionally exploiting temporal association between objects across frames. Novelly, to unify object localization and association in one architecture, we additionally learn an identifier for each instance and formulate it as a process of meta-learning [69]–[71], coupled with the training of general object detection. Specifically, at each training episode of meta-training stage, we first randomly choose a co-appearing object from two adjacent frames I_t and $I_{t+\tau}(\tau \in [-T, T])$ within temporal range T as the target object. Next, the perturbed boxes of this target object in frame I_t are set as the support set $\boldsymbol{\zeta}_S$ and that from frame $I_{t+\tau}$ are treated as the query set $\boldsymbol{\zeta}_Q$. Both of the support set $\boldsymbol{\zeta}_S$ and query set $\boldsymbol{\zeta}_Q$ are leveraged to perform a new task of instance association in current episode. The goal is to utilize the support set $\boldsymbol{\zeta}_S$ to learn a instance identifier that can accurately identify the target object in the query set. In this way, MINet is taught to learn instance identifiers on the fly, which can be well applied to up-coming frames. Both the meta-training process and the training of object detection can be jointly performed.

Besides, taking the inspiration from Memory Networks [72]–[76], we introduce a dynamic memory module at inference stage to record the detections of previous frames and further leverage the memorized detection history to rectify object detection in up-coming frame. Suppose the target frame to be detected at present is frame $\dot{I}_t$, and each detected object up-to frame $\dot{I}_{t-1}$ has been recorded as memorized object m_{t-1}^j in current memory $\mathcal{M}_{t-1}$ ($m_{t-1}^j \in \mathcal{M}_{t-1}$). Each sub-task of instance association is thus defined as identifying the same object m_{t-1}^j among the base detection boxes $\mathcal{P}_t$ within target frame $\dot{I}_t$. Here the corresponding instance identifier $\boldsymbol{\gamma}_{t-1}^j$ is online learnt from the support set $\boldsymbol{\zeta}_{S,t-1}^j$ which consists of the perturbation boxes of that target object m_{t-1}^j stored in memory $\mathcal{M}_{t-1}$. After associating each base detection box $p_t^i \in \mathcal{P}_t$ with the existing objects in memory, we can naturally enhance the base score $\boldsymbol{s}_t^i$ of p_t^i by aggregating the detection results of associated objects in memory. Furthermore, the dynamic memory module is in turn updated with the rectified detection results of frame $\dot{I}_t$.

B. Training With Meta-Learnt Instance Identifiers

One natural way to exploit temporal coherence across frames for video object detection is to perform box-level association. However, such direction commonly associates detected bounding boxes across consecutive frames via post-processing of linking or tracking, which might destroy the interaction between instance localization and association. Instead, we design a unified architecture to enable simultaneous object detection and instance association. This is achieved by meta-learning the instance identifiers that can be online adapted for instance association among frames.

The detailed architecture of our MINet for training is illustrated in Figure 2. In particular, we extend the base object detector by attaching an identify head that meta-learns instance-wise identifiers. The identify head is sibling to the detect head and both of them share the same backbone network (*i.e.*, ResNet-101). In identify head, each region feature is encoded into a 128-dimensional association embedding via two consecutive fully connected layers.

1) Support Set and Query Set: The meta training involves iterating through training episodes. We first present how to construct the support set and query set in each training episode.

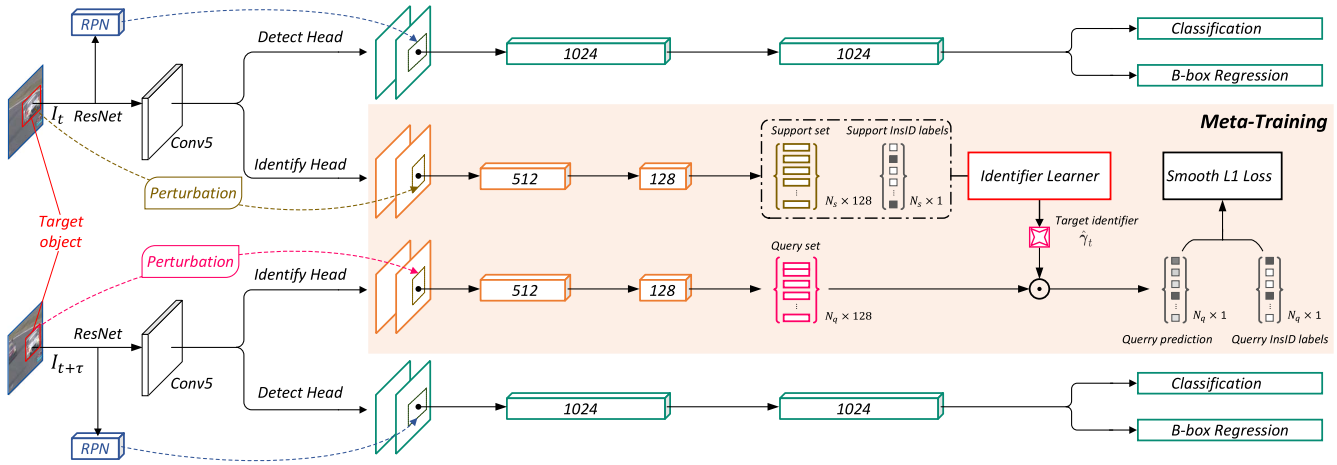


Fig. 2. An overview of our MINet with meta-training process at the training stage (better viewed in color). There are four main components in MINet: (1) a backbone network (*i.e.*, ResNet), (2) a region proposal network, (3) a detect head, and (4) an identity head. Given the input pair of adjacent frames $\{I_t, I_{t+\tau}\}$, the backbone network is leveraged to extract the feature map of each frame. Then, the detect head takes the region proposals generated by RPN and extracts the RoI feature over the feature maps out of the "Conv5" stage. The detect head performs classification and box regression as the standard paradigm of two-stage object detectors. The meta-training process is performed on the identity head. Specifically, a co-appearing object is first selected as the target object, and perturbed boxes are generated over the ground-truth box of the target object. The perturbed boxes from I_t and $I_{t+\tau}$ are taken as the support set and the query set, respectively. The identity head takes the association embeddings and InsID labels of the support set to learn an identifier, and then applies the identifier on the association embeddings of the query sets to generate the query prediction. The training objective of the meta-training process is computed by measuring the difference between the query prediction and query InsID labels with smooth L1 loss. Both the object detection process in detect head and the meta-training process in identify head can be jointly optimized at training.

Technically, we sample two frames I_t and $I_{t+\tau}$ ($\tau \in [-T, T]$) from a video, and ensure that there is at least one same instance appearing in both of them. Next, one of the co-appearing instances is randomly selected as the target object and a number of perturbation boxes are generated depending on the ground truth boxes of that target object in each frame as in [77]. Concretely, a positive threshold and a negative threshold is adopted to assign the instance identification label (InsID label) for these perturbed boxes. If the jaccard overlap between a perturbed box and the corresponding ground-truth box is larger than the positive threshold (0.7), this box is set as a positive sample (InsID label $= 1$). If the jaccard overlap is below the negative threshold (0.5), this box is set to be a negative sample (InsID label $= 0$). Otherwise, the box is ignored. We take a set of positive and negative perturbed boxes from I_t as support set $\boldsymbol{\zeta}_S$ and that from $I_{t+\tau}$ as query set $\boldsymbol{\zeta}_Q$. We summarize the detail of how to perform box perturbation in the Section III-D.

2) Learning Identifiers: Our MINet is taught to learn instance identifiers in each training episode with support set $\boldsymbol{\zeta}_S$, query set $\boldsymbol{\zeta}_Q$ and their corresponding InsID labels. Instead of learning the identifier via back-propagation which needs large computation, we follow [22], [70], [71], [78], [79] to teach our MINet adapting to unseen instances by predicting network parameters in forward propagation. Specifically, we meta-learn instance identifiers by adopting ridge regression as the optimization solver. Typically, ridge regression is a classical machine learning algorithm with closed-form solutions that predicts the regressor based on input data and labels. Compared with back-propagation that needs large computation, ridge regression which fully capitalizes on matrix-operation is more applicable for online adaptation and updating of identifiers. Specifically, the optimization objective for each instance identifier $\hat{\boldsymbol{\gamma}}$ is

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} \left\| y_s^n - \boldsymbol{w} \cdot \boldsymbol{a}_s^n \right\|_2 + \lambda \left\| \boldsymbol{w} \right\|_2, \tag{1}$$

where $\boldsymbol{a}_s^n \in \mathcal{A}_S$ is the association embedding of positive/negative box in support set and y_s^i indicates the corresponding InsID label. Note that $A_s \in \mathbb{R}^{N_s \times 128}$ is the set of association embeddings for N_s samples in support set $\boldsymbol{\zeta}_S$. $\|\boldsymbol{w}\|_2$ is the $L2$-norm regularization to avoid overfitting and λ is a scalar to balance the squared error and regularization. Equation 1 can be re-formulated with matrix operation, leading to the closed-form solution of $\hat{\boldsymbol{\gamma}}$:

$$\hat{\boldsymbol{\gamma}} = (A_s^T A_s + \lambda I)^{-1} A_s^T \boldsymbol{y}_s, \tag{2}$$

where $\boldsymbol{y}_s \in \mathbb{R}^{N_s \times 1}$ denotes the corresponding label set of A_s. I is the unitary matrix. The learnt identifier $\hat{\boldsymbol{\gamma}}$ is a 128-dimensional vector that acts as a linear classifier to distinguish positive boxes from negative ones.

3) Meta-Training Objective: Once the identifier $\hat{\boldsymbol{\gamma}}$ is obtained as in Equation 2, we apply it over boxes from the query set $\boldsymbol{\zeta}_Q$ to get the prediction of association scores. Specifically, for each association embedding $\boldsymbol{a}_q^j$ ($\boldsymbol{a}_q^j \in A_Q$) of query box, the estimated instance association score can be calculated by the inner product of $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{a}_q^j$. We perform meta-training with the objective to measure the smooth L_1 distance between the predicted instance association scores and InsID labels over the boxes from $\boldsymbol{\zeta}_Q$:

$$\mathcal{L}_{meta} = \frac{1}{N_Q} \sum_{j=1}^{N_Q} \text{smooth}_{L_1}(y_q^j - \hat{\boldsymbol{\gamma}} \cdot \boldsymbol{a}_q^j), \tag{3}$$

where N_Q is the number of query boxes in $\boldsymbol{\zeta}_Q$.

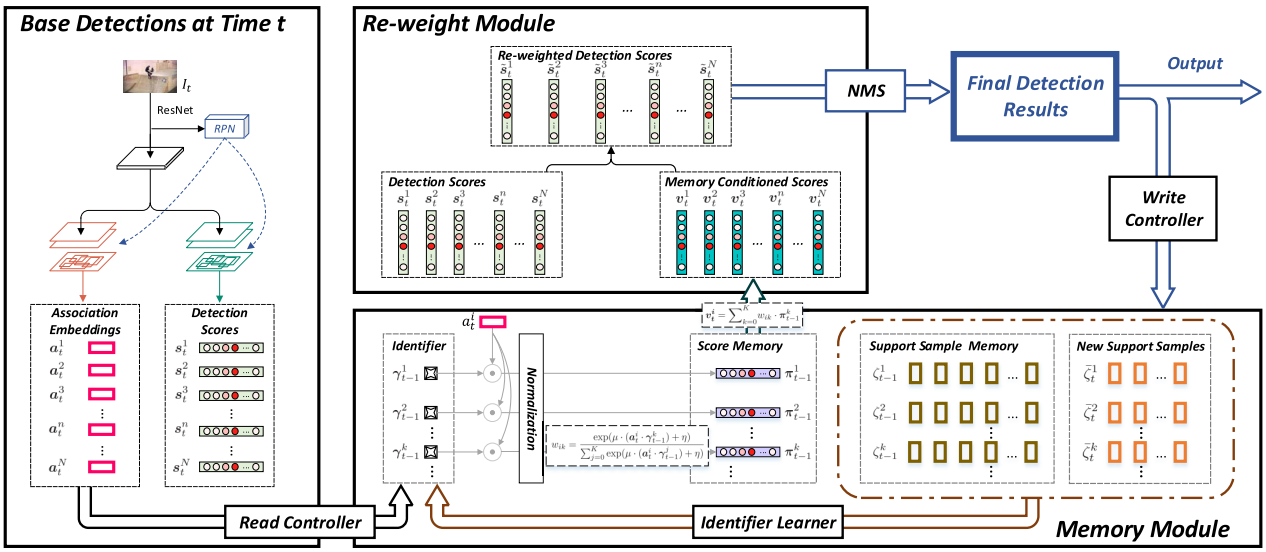


Fig. 3. An illustration of inference stage with dynamic memory module. We take frame I_t as current target frame to demonstrate the process of memory reading to refine detection results and memory writing/updating. For the base detections on frame I_t, we omit the 4-dimensional coordinates $\boldsymbol{b}_t^n$ for each detection box for simplicity.

4) Joint Optimization: In MINet, the detect head and identify head are integrated as a unified architecture, which are jointly optimized at training. The overall training objective is thus defined as the combination of the classification loss ($\mathcal{L}_{cls}$), the localization loss for region proposals ($\mathcal{L}_{loc}$), and meta training objective ($\mathcal{L}_{meta}$):

$$\mathcal{L} = \mathcal{L}_{meta} + \mathcal{L}_{cls} + \mathcal{L}_{loc} + \mathcal{L}_{rpn}. \tag{4}$$

Here $\mathcal{L}_{rpn}$ is the loss of region proposal network, which is also joint optimized as a part of our single-frame object detector, *i.e.*, Faster R-CNN.

C. Inference With Dynamic Memory Module

In this section, we devise a dynamic memory module to trace the detection results of existing objects in memory with meta-learnt instance identifiers, which further ameliorates object detection at inference stage. The detailed process of memory reading and writing/updating with regard to current target frame I_t is illustrated in Figure 3.

1) Base Detections: We firstly denote the outputs of detect head on frame I_t as $\mathcal{P}_t = \{p_t^i = (\boldsymbol{b}_t^n, \boldsymbol{s}_t^n, \boldsymbol{a}_t^n)\}_{n=1}^N$, which is the base detections of this frame. Specifically, $\boldsymbol{b}_t^n \in \mathbb{R}^4$ is the 4-dimensional coordinates of refined bounding box, $\boldsymbol{s}_t^n \in \mathbb{R}^{\mathcal{C}+1}$ denotes the $(\mathcal{C}+1)$-dimensional detection score, and $\boldsymbol{a}_t^n$ is the association embedding out of the identify head. Here the first dimension of $\boldsymbol{s}_t^n$ indicates the background confidence and the others are confidences for the $\mathcal{C}$ foreground classes.

2) Memory: The recorded memory up-to frame I_{t-1} is defined as $\mathcal{M}_{t-1} = \{(\boldsymbol{\pi}_{t-1}^k, \boldsymbol{\gamma}_{t-1}^k, \boldsymbol{\zeta}_{t-1}^k)\}_{k=1}^K$. Each memory slot consists of the memory score $\boldsymbol{\pi}_{t-1}^k$, an identifier $\boldsymbol{\gamma}_{t-1}^k$ and a support sample buffer $\boldsymbol{\zeta}_{t-1}^k$. For each $m_{t-1}^k \in \mathcal{M}_{t-1}$, $\boldsymbol{\pi}_{t-1}^k$ is a $(\mathcal{C}+1)$-dimensional vector that denotes the memorized detection score of the k-th instance. $\boldsymbol{\gamma}_{t-1}^k$ denotes the instance identifier meta-learnt at time $t-1$. $\boldsymbol{\zeta}_{t-1}^k$ is a sample buffer with fixed capacity to store support samples from preceding T frames. Note that an invalid memory slot m_{t-1}^0 is further added to deal with newly appearing object. The detection score $\boldsymbol{\pi}_{t-1}^0$ of m_{t-1}^0 is set as $\frac{1}{\mathcal{C}+1}$ for each class, thus it provides no prior information for classification.

3) Read Controller: Different from existing memory networks that access memory via a read controller in key-value structure, we leverage meta-learnt identifiers to control memory reading. Technically, for the i-th proposal $p_t^i \in \mathcal{P}_t$, we first measure the association score between p_t^i and each memorized object m_{t-1}^k as the dot-product between identifier $\boldsymbol{\gamma}_{t-1}^k$ and association embedding $\boldsymbol{a}_t^i$, which is further normalized as:

$$w_{ik} = \frac{\exp(\mu \cdot (\boldsymbol{a}_t^i \cdot \boldsymbol{\gamma}_{t-1}^k) + \eta)}{\sum_{j=0}^K \exp(\mu \cdot (\boldsymbol{a}_t^i \cdot \boldsymbol{\gamma}_{t-1}^j) + \eta)}, \tag{5}$$

where μ and η are scale and shift factor, and K denotes the number of valid slots. Then, we retrieve the memory conditioned score $\boldsymbol{v}_t^i$ by aggregating all memory scores weighted with association scores:

$$\boldsymbol{v}_t^i = \sum\nolimits_{k=0}^K w_{ik} \cdot \boldsymbol{\pi}_{t-1}^k. \tag{6}$$

4) Detection Score Re-Weighting: Since $\boldsymbol{s}_t^i$ is predicted on the appearance feature at present and $\boldsymbol{v}_t^i$ is aggregated by the temporal association with previous frames, they are naturally complementary to each other. Hence we can obtain the re-weighted detection score $\tilde{\boldsymbol{s}}_t^i$ by mixing them as:

$$\tilde{s}_t^{i,c} = \frac{s_t^{i,c} \cdot v_t^{i,c}}{\sum_{q=0}^C s_t^{i,q} \cdot v_t^{i,q}}, \tag{7}$$

where $s_t^{i,c} \in \boldsymbol{s}_t^i$ and $v_t^{i,c} \in \boldsymbol{v}_t^i$ is the detection/memory conditioned score of the c-th class, and $\tilde{s}_t^{i,c} \in \tilde{\boldsymbol{s}}_t$ indicates the re-weighted score. Such re-weighting operation encourages the detection of current frame to be consistent with previous ones. Then, class-wise non-maximum suppression (NMS) is applied with re-weighted score to get the final prediction $\mathcal{Z}_t$.

5) Write Controller: Memory $\mathcal{M}_{t-1}$ is further updated by detection results $\mathcal{Z}_t$ after NMS. The updating strategy of the write controller consists of three steps: box-to-memory matching, memory score updating, and updated identifier learning. In particular, we first assign $z_t^i \in \mathcal{Z}_t$ and $m_{t-1}^k \in \mathcal{M}_{t-1}$ as the vertices on two sides of a bipartite graph. Each edge in this graph represents the association score. Hungarian algorithm is then exploited to produce matching pairs. The pairs with low association score are discarded. Taking the matching pair (z_t^p, m_{t-1}^q) as an example, we update the memory score $\boldsymbol{\pi}_{t-1}^q$ as:

$$\boldsymbol{\pi}_t^q = \frac{\boldsymbol{s}_t^p + \sigma \cdot f_{\text{num}}(m_{t-1}^q) \cdot \boldsymbol{\pi}_{t-1}^q}{1 + \sigma \cdot f_{\text{num}}(m_{t-1}^q)}, \tag{8}$$

where $\boldsymbol{s}_t^p$ is the base detection score of z_t^p, $f_{\text{num}}(m_{t-1}^q)$ indicates the number of previous matched boxes for m_{t-1}^j, and σ is the decay factor. The last step is to update identifier $\boldsymbol{\gamma}_{t-1}^q$. We generate new support samples based on z_t^p and add them into the sample buffer $\boldsymbol{\zeta}_{t-1}^q$ to get $\boldsymbol{\zeta}_t^q$. Then, in the identifier learner, $\boldsymbol{\zeta}_t^q$ is taken as support set to learn the new identifier $\boldsymbol{\zeta}_t^q$ as in Equation 2. Note that the memory is empty at the beginning of each video and we initialize the memory $\mathcal{M}_1$ with base detection results of the first frame.

D. Meta-Learning Sample Generation

In this section, we detail the sampling process for meta-training and meta-testing. Typically, in meta-training, samples from both support and query set are generated by box perturbation over ground truth boxes. In meta-testing, the support samples are obtained from the perturbation over the rectified detection boxes, while the query samples are the base detections of the up-coming frame.

1) Meta-Training: In meta-training, support samples and query samples are generated based on the ground truth bounding box of target object like [35], [77], [80]. Firstly, we represent a ground truth box with the center point coordinates, width and height of this box as $[x_c, y_c, w, h]$. And then, we add gaussian noise to these four elements with variance randomly chosen from $\{0.01, 0.05, 0.1, 0.2, 0.3\}$ to generate perturbation boxes. The support set is composed of perturbation boxes, and the InsID label are assigned according to the jaccard overlaps with $[x_c, y_c, w, h]$.

2) Meta-Testing: In meta-testing, the positive and negative samples of support set are generated with different strategies. Similar to the ground truth box, we denote the rectified detection box as $[\dot{x}_c, \dot{y}_c, \dot{w}, \dot{h}]$. On the one hand, we translate the matched box to $[\dot{x}_c + \delta_x, \dot{y}_c + \delta_y, \dot{w}, \dot{h}]$, where $\delta_x \in \{\dot{x}_c + k \cdot \dot{w}/12\}_{k=-4}^4$ and $\delta_y \in \{\dot{y}_c + k \cdot \dot{h}/12\}_{k=-4}^4$ to generate positive sample candidates. Then, we filter them according to the jaccard overlaps with $[\dot{x}_c, \dot{y}_c, \dot{w}, \dot{h}]$ and select no more than 32 boxes as positive support samples. On the other hand, we re-utilize the region proposals in this frame and sample 96 negative support samples from them. Since we have already extracted the association embeddings for all the proposals for memory reading, such a strategy reduces the computation cost for online support samples feature extraction.

IV. Experiments

A. Dataset

We evaluate our MINet on ImageNet object detection from video (VID) dataset. The publicly accessible part of this dataset contains 30 basic-level categories with 3,862 video snippets for training and 555 snippets for validation. A subset of ImageNet object detection (DET) dataset with intersected 30 classes is also utilized for training. We utilize the sampled frames/images released by [18], [81] with 15 frames from each snippet in ImageNet VID dataset and about 2,000 images per each class from ImageNet DET dataset. For evaluation, we follow the widely adopted protocol to report the performance on validation set under the measurement of mean average precision (mAP).

B. Implementation Details

1) Network Architecture: We utilize ResNet-101 as the backbone network in our MINet. The $conv5$ stage is modified to enlarge the receptive field by decreasing the stride from 2 to 1 and increasing the dilation ratio of each 3×3 convolution layers in this stage from 1 to 2. The region proposal network (RPN) is built on the top of $conv4$ stage for region proposal generation. Anchors of 4 scales $\{512^2, 256^2, 128^2, 64^2\}$ and 3 aspect ratios $\{2:1, 1:1, 1:2\}$ are assigned to each grid cell on the feature map. The detect head and identify head is applied on the output of $conv5$ stage. Since the dimension of $conv5$ feature is 2048 that will results in large computation overhead for region feature extraction, we attach additional conv layers at the beginning of each head to reduce the feature dimension from 2048 to 256. In detect head, we follow [82] to utilize two large separable conv layers [83], [84] before ROI feature extraction. The kernel size of these conv layers is set as 15. Two 1024-dim fully connected layers are exploited in subnet before classification and bounding box regression. In identify head, two 3×3 conv layers are applied on $conv5$ feature to simultaneously reduce the feature dimension and decouple instance identification from object detection. We encode each region into a 128-dim association feature embedding after an internal 512-dim fully connected layer.

2) Training: Our MINet is trained end-to-end over 4 Tesla Titan V GPUs with SGD optimizer. The weight decay is set as 0.0001 and the momentum is 0.9. We assign one mini-batch to each GPU and thus the total batch size is 4. In each mini-batch, two adjacent frames are randomly sampled from a video within temporal range $T = 30$. Typically, it is ensured to be at least one co-appearing object in each pair of frames. We resize the image to a shorter side no more than 600 pixels and longer side no more than 1000 pixels. The training process takes $130K$ iterations in total. The learning rate is set as 0.001 in the initial $80K$ iterations of training and divided by 10 at the $80K$ and $120K$ iterations. In detect head, we randomly sample 128 ROIs for training, and the ratio between positive and negative ROIs is set up to $1:3$. In identify head, the number of boxes in each support set is 128, and the positive-to-negative ratio is set as $1:3$. Besides, the regularization parameter λ is set as 0.1.

3) Inference: For each frame, 300 region proposals output from RPN with NMS of threshold 0.7 are fed into

TABLE II

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON IMAGENET VID VALIDATION SET. ALL OF THE METHODS EXPLOIT RESNET-101 AS THEIR BACKBONE NETWORK, EXCEPT FOR THP AND STSN THAT USE DEFORMABLE CONVOLUTIONAL [7] LAYERS IN THEIR BACKBONE. THE ABBREVIATIONS "FA" AND "BA" REPRESENT THE FEATURE AGGREGATION PARADIGM AND THE BOX ASSOCIATION PARADIGM, RESPECTIVELY. MINET$^{\triangle}$ INDICATES THAT WE MERGE THE DETECTION RESULTS OF BOTH FORWARD AND BACKWARD PASSES FOR EACH VIDEO

Methods	Backbone	Paradigm	Online	Offline	mAP (%)	Runtime (ms)
FGFA [18]	ResNet-101	FA	✓		76.3	330
MANet [20]	ResNet-101	FA	✓		78.1	202
THP [49]	ResNet-101 + DCN [7]	FA	✓		78.6	-
STSN [45]	ResNet-101 + DCN [7]	FA	✓		78.9	-
LWDN [85]	ResNet-101	FA	✓		76.3	-
PSLA [46]	ResNet-101	FA	✓		77.1	32
OGEM [47]	ResNet-101	FA	✓		79.3	112
SELSA [54]	ResNet-101	FA		✓	80.3	-
RDN [50]	ResNet-101	FA	✓		81.8	115
MEGA [55]	ResNet-101	FA		✓	82.9	-
Faster R-CNN + Seq-NMS [16]	ResNet-101	BA		✓	77.5	-
Faster R-CNN + Tube-Link [86]	ResNet-101	BA		✓	77.1	-
D&T-online [15]	ResNet-101	BA	✓		78.7	141
D&T($\tau = 1$) [15]	ResNet-101	BA		✓	79.8	-
ST-lattice [52]	ResNet-101	BA		✓	79.6	-
FGFA [18] + Seq-NMS [16]	ResNet-101	FA + BA		✓	78.4	-
MANet [20] + Seq-NMS [16]	ResNet-101	FA + BA		✓	80.3	-
STMN [19] + Seq-NMS [16]	ResNet-101	FA + BA		✓	80.5	-
SELSA [54] + Seq-NMS [16]	ResNet-101	FA + BA		✓	80.5	-
PSLA [46] + Seq-NMS [16]	ResNet-101	FA + BA		✓	78.6	-
MINet	ResNet-101	BA	✓		**80.2**	133
MINet$^{\triangle}$	ResNet-101	BA		✓	**80.6**	-

detect/identify head to get the base detection boxes and association embeddings. We only take base detection with foreground score higher than 0.1 to initialize and update memory module. For each memory cell $m^i \in \mathcal{M}$, the capacity of sample buffer $\boldsymbol{\zeta}^i$ is set to record samples from at most 30 preceding matched boxes. In read controller, we manually set the association score between m_t^0 and every detection box to be 0.5. The scale and shift factor (*i.e.*, μ and η) are set as 20 and -10 empirically. NMS with threshold of 0.45 is adopted after re-weighting detection score to get the final prediction. Decay factor σ is set as 0.95 in Equation 8. For each cell to be updated, 128 new support samples are generated and added to the corresponding sample buffer.

C. Comparison With State-of-the-Art Methods

We compare our MINet with several state-of-the-art video object detection methods on ImageNet VID validation set and summarize the performance in Table II. Note that Faster R-CNN+ [16] and Faster R-CNN+ [86] are our re-implementations by integrating Seq-NMS [16] and tube linking algorithms [86] into Faster R-CNN, respectively. In general, all state-of-the-art approaches here can be grouped into two categories, *i.e.*, feature aggregation based approaches and algorithms with box-level association. Our MINet belongs to the latter one.

Among the box-level association methods, the former four represent detection boxes as vertices in linkage graphs and formulate the tubelet linking as an optimal path finding problem. In [16] and [86], linkage graphs are constructed only based on the geometric information (*i.e.*, the coordinates of bounding boxes) and detection score output from object detectors. D&T($\tau = 1$) achieves better performance than [16] and [86] by involving tracklets between consecutive frames into linkage graph, which benefits association of moving objects. After box-level association with propagation relations, ST-lattice [52] additionally leverages R-CNN like classifier to re-classify each bounding box, which further boosts the performance. On the other side, feature aggregation based models encourage temporal coherence among detection results of consecutive frames by leveraging extra input (*i.e.*, optical flow in FGFA [18], THP [49] and MANet [20]) or capitalizing on memory modules (*i.e.*, PSLA [46], OGEM [47]). The overall results with the same backbone network demonstrate that our proposed MINet by integrating object detection and association in a unified architecture with meta-learnt identifiers exhibits better performance than all the other methods. Moreover, the further performance improvement is attained when we combine the detection results from forward and backward passes.

Generally, box-level association based algorithms commonly take sequence-level information to find the optimal tubelets and re-score boxes in each tube, which is taken as post-processing to refine predictions in each single frame. When applying the box-level association based methods in online setting, one natural solution is to conduct box linking

TABLE III
ABLATION STUDY ON EACH COMPONENT OF OUR MINET. WE BUILD OUR MODEL ON THE TOP OF RESNET-101 AND TAKE FASTER R-CNN AS THE BASE OBJECT DETECTOR

Methods	Meta-training	Memory module	mAP (%)
(a)			75.6
(b)	✓		$77.6_{\uparrow 2.0}$
(c)		✓	$77.5_{\uparrow 1.9}$
(d)	✓	✓	$\mathbf{80.2}_{\uparrow \mathbf{4.6}}$

only based on previous detections, which inevitably results in performance drop. For instance, D&T (online) is an online version of D&T($\tau = 1$) with inferior performance. In contrast to algorithms that associate boxes with linkage graph (*i.e.* [16], [86] and D&T [15]) or tracking (*i.e.* ST-lattice [52]), our MINet enables online instance association by meta-learning instance identifiers, which obtains better performance without any post-processing.

Although our method has not achieved the best performance among all the competitors, the proposed MINet shows its benefits for the following aspects. Compared to the FA methods [54], [55] that rely on temporal coherence of the whole video for prediction, our MINet is an online and causal system. Typically, MINet only exploits the inputs of the past time and the current time, which is more applicable for video applications. Besides, it is a tendency for the community to jointly consider the problem of video objection and multiple object tracking, since they are closely related. Although RDN [50] is also an online algorithm performing in the box level, it leverages the relation between region proposals of multi-frames in an implicit manner. In an other word, RDN can improve the detection performance, but cannot generate tracklets. On the contrary, our MINet explicitly associates the per-frame detections to the memory slot, and thus generates tracklets as a by-product of object detection. In light of MINet's capability to explicitly perform box association, our method exhibits great potential to bridge these two topics. Among the box association methods, our MINet shows the best mean average precision. Particularly, in comparison with the method which can also perform online inference (*i.e.*, D&T-online), our method achieves 1.5% absolutely improvements. Moreover, our MINet can also be applied to the feature aggregation methods for further improvements. We have conducted extended experiment with the most representative FGFA in Section IV-H. Moreover, our MINet conveys an elegant view of exploiting the meta knowledge to learn the parameters of instance identifiers on the fly. This is a problem that has not been fully understood in the previous literatures of this field.

D. Ablation Study

We investigate the influences of each component in our MINet for video object detection, as summarized in Table III. All of the models are built on the backbone of ResNet-101 and Faster R-CNN is adopted as the base object detector.

1) Method (a): Is the still image object detector (*i.e.*, Faster R-CNN), which directly performs object detection over each single frame but ignores the temporal association between instances across frames. The mAP performance of (a) is 75.6%, which serves as a strong baseline to evaluate relative improvements of each components in our methods.

2) Method (b): Exploits the meta-training process of instance identifiers and integrates it into the training process of (a), which makes the absolute improvement of mAP score over (a) by 2.0%. The meta-training influences the detection performance of our model in two aspects. On the one hand, the meta-training process benefits feature learning of backbone networks by composing to multi-task training. On the other hand, since the support set to learn the target instance identifier and the query set to be measured on are generated from different frames, the meta-training effects as a Dconsistent constraint over region features across frames.

3) Method (c): Equips (a) with the dynamic memory module at the inference stage, but without involving meta-training. Instead of exploiting the association embeddings encoded in identify head, we utilize the $fc6$ feature from the detect head for online identifier learning and box-to-memory association. The performance of (c) is 77.5%, which is higher than still image detector (a) but sub-optimal compared with our overall architecture in (d). This result further validates the effectiveness of meta-training in our MINet.

4) Method (d): Indicates the integral architecture of our MINet, which increases the mAP to 80.2% by capitalizing on both meta-training in training phase and dynamic memory module in inference phase. This result demonstrates the advantage of utilizing meta-leant instance identifiers to associate detection boxes with existing instances and leveraging memorized score to refine single frame detections.

E. Qualitative Analysis

1) Robustness of Object Detection: Four video examples of detection results on ImageNet VID validation set by still-image detector (*i.e* Faster R-CNN) and our MINet are displayed in Figure 4. We showcases detection boxes with confidence higher than 0.2 on three frames from each video. The red boxes are true-positive detections and blue boxes indicates prediction with wrong categories or low localization quality. MINet consistently shows better detection results than the still-image object detector. For example, in the first video, the still-image detector fails to detect the "bears" when the frames zoom out. In contrast, by associating detection boxes in these frame with memorized objects, our MINet can detect these bears correctly. Furthermore, in the line chart, we depict the score of true-positive detections across frames (if there is no true-positive in this frame, the sore is set as 0). Obviously, detection scores of the still-image detector suffers from large fluctuations across frames. In contrast, by leveraging meta-learnt instance identifiers to associate per-frame detections to objects in memory, the detection results of MINet is significantly more consistent and robust.

2) Capability of Box Association: In this experiment, we choose one of the ground truth boxes from the beginning frame and region proposals from up-coming frames to demonstrate the capability of box association of our MINet.

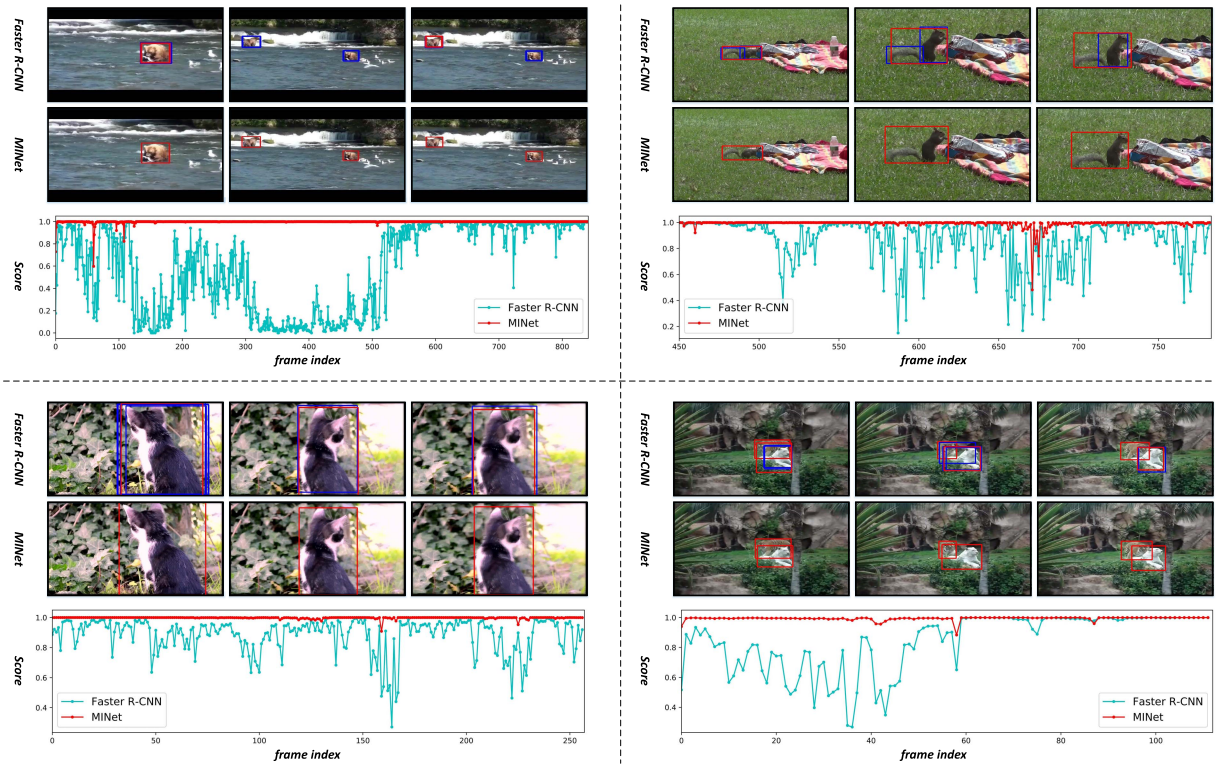


Fig. 4. Four video examples of detection results by the still-image object detector (*i.e.* Faster RCNN) and our MINet. For each video, we depict detection results from three frames and details the detection score of true-positive predictions in the line chart. Typically, the red boxes denote true-positive detections and blue boxes indicate prediction with wrong categories or low localization quality.

Specifically, we first take the ground truth boxes from the beginning frame of each sequence to initialize the dynamic memory module. Then, for the up-coming frames, we take the RPN for region proposal generation and extract the association embedding for each region proposal. Box association score is obtained by performing dot-product between the meta-learnt identifier and the association embedding of each proposal. Finally, we choose the proposal with highest association score as the associated box of the chosen instance. We depict the results in Figure 5, where the red boxes on the left are the chosen ground truth boxes and the cyan boxes on the right are the associated boxes of the up-coming frames. As shown in this figure, MINet can make correct association prediction with the meta-learnt identifier, which further demonstrates the effectiveness of our identify head.

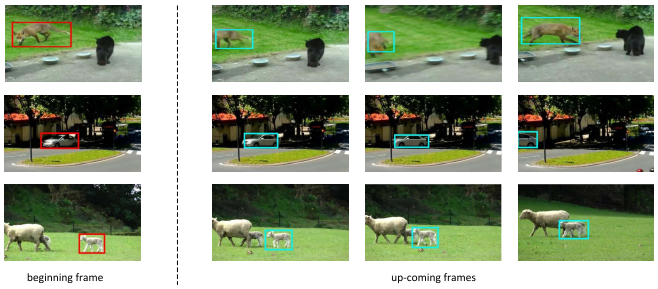


Fig. 5. Three video example to demonstrate the capability of box association of our MINet. The chosen ground truth boxes (in red) of each video are on the left, while the associated box (in cyan) from the up-coming frames are on the right.

TABLE IV

PERFORMANCE OF ADDING SEQUENCE-LEVEL BOX ASSOCIATION METHODS TO THE DETECTION RESULTS OF OUR MINET

Ours	+ [16]	+ [86]	mAP (%)
✓			80.2
✓	✓		80.4
✓		✓	80.2

F. Experimental Analysis

1) Extra Post-Processing: In Table IV, we additionally add two widely adopted sequence-level box association techniques (*i.e.* [16] and [86]) to investigate whether the dynamic memory module in our MINet is capable enough to capture the temporal coherence among frames. When equipped with [16] or [86], there is not distinct improvement over our online inference, which consolidates that the effectiveness of our inference with dynamic memory.

2) Effect of Decay Factor in Write Controller: Table V compares the performance by varying the score decay factor σ in our write controller. When $\sigma = 1.0$, the memory score is updated as the average detection scores of matched objects up-to this moment. For the extreme choice of $\sigma = 0$, the detection score of previous matched objects will all be forgotten in updating, and only the detection score at present is recorded. We set $\sigma = 0.95$ by default, since it make a good balance between previous results and the present one.

TABLE V
PERFORMANCE COMPARISON WITH DIFFERENT MEMORY SCORE DECAY FACTOR σ IN OUR WRITE CONTROLLER. (* INDICATES DEFAULT SETTING)

σ	1.0	0.95*	0.90	0.80	0.70	0.50	0.25	0.0
mAP (%)	80.22	80.24	80.17	80.07	80.07	79.95	79.83	78.4

Fig. 6. Failure cases in the validation set of ImageNet VID dataset (better viewed in color). The red boxes illustrate the true-positive detections. The blue box means that the classification prediction is right, but the IoU between the predicted box and the ground-truth one are less than 0.5. The purple boxes indicates the detections are wrongly categorized.

G. Failure Cases Study

In Figure 6, we showcase two failure cases of the proposed MINet. At the beginning of sequence (a), only a small part of a snake appears in the shot, so that the identifier learns to associate this part to the corresponding memory slot. When a larger part is captured in the shot, the memory slot will promote the score of the proposal that contains a similar small part as the previous frames, leading to false alarm errors. Our MINet is capable to alleviate this issue by itself. Specifically, since the correct prediction (the red box in the third column) will not be associated to the existing memory slots, a new memory slot will be initialized with this prediction. Thus, in the following frames, the classification confidence of a larger part of this snake will also be promoted. In case (b), a monkey hides behind the leaves. This monkey is wrongly categorized as "bear" and "red panda" across the whole video. Our MINet cannot mitigate the classification errors in an up-coming frame if none of the previous frames have made true-positive predictions. The sequence of case (b) is an extreme hard case of ImageNet VID dataset. Even we human beings cannot recognize the monkey in these frames. The existing of such hard cases accounts for the reason why the performance on this dataset has become almost saturated.

H. Generalization Capability

To further validate the capability of our MINet to perform as a plug-in module, we conduct an experiment to combine the proposed method with the most representative feature aggregation method (*i.e.*, FGFA [18]). Typically, FGFA is a feature aggregation method that leverages optical flow as guidance to take motion compensation on feature maps of adjacent frames and aggregates them into the target one. In this experiment, the models are only trained on the ImageNet VID dataset. The performance of the FGFA baseline on the validation set of ImageNet VID dataset is 72.6% mAP. When adding the identify head and further performing memory-based inference as in MINet, the performance is boosted to 74.1% mAP, achieving 1.5% absolute improvement. Note that when comparing FGFA + MINet to the run of integrating MINet into single-frame detector (*i.e.*, Faster R-CNN [13]), we find that the meta-training process has not caused extra improvement over the FGFA baseline. One possible reason accounts for this phenomenon is that the meta-training process acts as a consistent constraint over features across frames (as analyzed in Section IV-D), while FGFA has exploited the temporal consistence through multi-frame feature aggregation to some extend.

V. CONCLUSION

We have presented Meta-Learnt Instance Identifier Networks (MINet), which enables a unified architecture of object detection and association in videos without any auxiliary input or post-processing. Particularly, we study the problem from the viewpoint of meta-learning instance identifiers to recognize identical instances across frames in a single forward-pass. To verify our claim, we teach MINet to learn instance identifiers during meta-training stage, whose parameters are produced on the fly conditioned on the detection results of previous frames. The meta-learnt instance identifiers are enforced to perform well when applied to up-coming frames. Moreover, at inference, a dynamic memory is further exploited to manage the history of the detected instances in pervious frames. The memorized objects are leveraged to enhance the predictions of detected instances in the up-coming frame via box-to-memory association. Experiments conducted on ImageNet VID demonstrate the efficacy of MINet. Performance improvements are clearly observed when comparing to existing techniques.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[5] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.
[6] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. NIPS*, 2016, pp. 379–387.
[7] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. ICCV*, 2017, pp. 764–773.
[8] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440–1448.
[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 2117–2125.
[12] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. CVPR*, Jun. 2019, pp. 821–830.
[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[14] Q. Cai, Y. Pan, Y. Wang, J. Liu, T. Yao, and T. Mei, "Learning a unified sample weighting network for object detection," in *Proc. CVPR*, Jun. 2020, pp. 14173–14182.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proc. ICCV*, 2017, pp. 3038–3046.

[16] W. Han *et al.*, "Seq-NMS for video object detection," 2016, *arXiv:1602.08465*. [Online]. Available: http://arxiv.org/abs/1602.08465

[17] K. Kang *et al.*, "Object detection in videos with tubelet proposal networks," in *Proc. CVPR*, Jul. 2017, pp. 727–735.

[18] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. CVPR*, Oct. 2017, pp. 408–417.

[19] F. Xiao and Y. Jae Lee, "Video object detection with an aligned spatial-temporal memory," in *Proc. ECCV*, 2018, pp. 485–501.

[20] S. Wang, Y. Zhou, J. Yan, and Z. Deng, "Fully motion-aware network for video object detection," in *Proc. ECCV*, 2018, pp. 542–557.

[21] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.

[22] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *Proc. ICLR*, 2019, pp. 1–15.

[23] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (CVPR)*, Oct. 2019, pp. 9925–9934.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Jun. 2018, pp. 7132–7141.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988.

[26] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.

[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. CVPR*, Jul. 2017, pp. 7263–7271.

[29] H. Lee, S. Eum, and H. Kwon, "ME R-CNN: Multi-expert R-CNN for object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 1030–1044, 2020.

[30] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask SSD: An effective single-stage approach to object instance segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 2078–2093, 2020.

[31] Y.-L. Li and S. Wang, "HAR-Net: Joint learning of hybrid attention for single-stage object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3092–3103, 2020.

[32] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[33] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. CVPR*, Jun. 2018, pp. 6154–6162.

[34] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. CVPR*, Jun. 2019, pp. 7036–7045.

[35] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. ECCV*, 2018, pp. 784–799.

[36] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. CVPR*, Jun. 2019, pp. 2239–2247.

[37] Y. Pan, T. Yao, Y. Li, C.-W. Ngo, and T. Mei, "Exploring category-agnostic clusters for open-set domain adaptation," in *Proc. CVPR*, Jun. 2020, pp. 13867–13875.

[38] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. CVPR*, Jun. 2019, pp. 11457–11466.

[39] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. ICCV*, 2019, pp. 480–490.

[40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[41] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. ECCV*, 2018, pp. 734–750.

[42] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. ICCV*, Oct. 2019, pp. 9627–9636.

[43] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. CVPR*, Jun. 2019, pp. 850–859.

[44] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. CVPR*, Jun. 2019, pp. 840–849.

[45] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proc. ECCV*, 2018, pp. 331–346.

[46] C. Guo *et al.*, "Progressive sparse local attention for video object detection," in *Proc. ICCV*, 2019, pp. 3909–3918.

[47] H. Deng *et al.*, "Object guided external memory network for video object detection," in *Proc. ICCV*, 2019, pp. 6678–6687.

[48] M. Zhu and M. Liu, "Mobile video object detection with temporally-aware feature maps," in *Proc. CVPR*, Jun. 2018, pp. 5686–5695.

[49] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proc. CVPR*, Jun. 2018, pp. 7210–7218.

[50] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Relation distillation networks for video object detection," in *Proc. ICCV*, Oct. 2019, pp. 7023–7032.

[51] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Single shot video object detector," *IEEE Trans. Multimedia*, vol. 23, pp. 846–858, 2021.

[52] K. Chen *et al.*, "Optimizing video object detection via a scale-time lattice," in *Proc. CVPR*, Jun. 2018, pp. 7814–7823.

[53] K. Kang *et al.*, "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2896–2907, Oct. 2018.

[54] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proc. ICCV*, 2019, pp. 9217–9225.

[55] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proc. CVPR*, Jun. 2020, pp. 10337–10346.

[56] Z. Zhang, D. Cheng, X. Zhu, S. Lin, and J. Dai, "Integrated object detection and tracking with tracklet-conditioned detection," 2018, *arXiv:1811.11167*. [Online]. Available: http://arxiv.org/abs/1811.11167

[57] E. Park and A. C. Berg, "Meta-tracker: Fast and robust online adaptation for visual object trackers," in *Proc. ECCV*, 2018, pp. 569–585.

[58] G. Wang, C. Luo, X. Sun, Z. Xiong, and W. Zeng, "Tracking by instance detection: A meta-learning approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6288–6297.

[59] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3623–3632.

[60] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9236–9245.

[61] T. Yang and A. B. Chan, "Learning dynamic memory networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 152–167.

[62] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. ICCV*, 2019, pp. 9226–9235.

[63] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Van Gool, "Video object segmentation with episodic graph memory networks," 2020, *arXiv:2007.07020*. [Online]. Available: http://arxiv.org/abs/2007.07020

[64] J. Miao, Y. Wei, and Y. Yang, "Memory aggregation networks for efficient interactive video object segmentation," in *Proc. CVPR*, Jun. 2020, pp. 10366–10375.

[65] Y. Chen, Y. Pan, T. Yao, X. Tian, and T. Mei, "Mocycle-GAN: Unpaired video-to-video translation," in *Proc. ACM Multimedia*, 2019, pp. 647–655.

[66] Y. Pan, Y. Li, T. Yao, T. Mei, H. Li, and Y. Rui, "Learning deep intrinsic video representation by exploring temporal coherence and graph structure," in *Proc. IJCAI*, 2016, pp. 3832–3838.

[67] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. CVPR*, Jun. 2016, pp. 4594–4602.

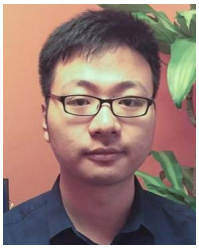
[68] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proc. CVPR*, Jun. 2019, pp. 284–293.

[69] M. Andrychowicz *et al.*, "Learning to learn by gradient descent by gradient descent," in *Proc. NIPS*, 2016, pp. 3981–3989.

[70] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. NIPS*, 2016, pp. 523–531.

[71] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. ICML*, 2017, pp. 2554–2563.
[72] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," 2016, *arXiv:1606.03126*. [Online]. Available: http://arxiv.org/abs/1606.03126
[73] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. ICML*, 2016, pp. 1842–1850.
[74] S. Sukhbaatar *et al.*, "End-to-end memory networks," in *Proc. NIPS*, 2015, pp. 1–11.
[75] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *Proc. ICLR*, 2014, pp. 1–15.
[76] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proc. CVPR*, Jun. 2018, pp. 4080–4088.
[77] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. CVPR*, Jun. 2019, pp. 4660–4669.
[78] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proc. NIPS*, 2017, pp. 1–12.
[79] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks," *Neural Comput.*, vol. 4, no. 1, pp. 131–139, Jan. 1992.
[80] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. CVPR*, Jun. 2016, pp. 4293–4302.
[81] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. CVPR*, Jul. 2017, pp. 2349–2358.
[82] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: In defense of two-stage object detector," 2017, *arXiv:1711.07264*. [Online]. Available: http://arxiv.org/abs/1711.07264
[83] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. CVPR*, Jul. 2017, pp. 4353–4361.
[84] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, Jun. 2016, pp. 2818–2826.
[85] Z. Jiang, P. Gao, C. Guo, Q. Zhang, S. Xiang, and C. Pan, "Video object detection with locally-weighted deformable neighbors," in *Proc. AAAI*, 2019, pp. 8529–8536.
[86] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. CVPR*, Jun. 2015, pp. 759–768.

Jiajun Deng (Graduate Student Member, IEEE) received the B.E. degree in electrical engineering and information science and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China in 2016 and 2021, respectively. His research interests include computer vision, 3D scene understanding, and vision-language understanding.



Yingwei Pan received the Ph.D. degree in electrical engineering from the University of Science and Technology of China in 2018. He was a Core Designer of top-performing multimedia analytic systems in worldwide competitions, such as COCO Image Captioning, Visual Domain Adaptation Challenge 2018, ActivityNet Dense-Captioning Events in Videos Challenge 2017, and MSR-Bing Image Retrieval Challenge 2014 and 2013. He is currently a Researcher with Vision and Multimedia Laboratory, JD AI Research, Beijing, China. His research interests include vision and language, domain adaptation, and large-scale visual search.

Ting Yao (Member, IEEE) is currently the Principal Researcher with Vision and Multimedia Laboratory, JD AI Research, Beijing, China. Prior to joining JD.com, he was a Researcher with Microsoft Research Asia, Beijing. He is the Principal Designer of several top-performing multimedia analytic systems in international benchmark competitions, such as ActivityNet Large Scale Activity Recognition Challenge for the period 2019–2016, Visual Domain Adaptation Challenge for the period 2019–2017, and COCO Image Captioning Challenge. He is a Leading Organizer of MSR Video to Language Challenge in ACM Multimedia 2017 and 2016, and built MSR-VTT, a large-scale video to text dataset that is widely used worldwide. His research interests include video understanding, vision and language, and deep learning. His works have led to many awards, including the ACM SIGMM Outstanding Ph.D. Thesis Award 2015, the ACM SIGMM Rising Star Award 2019, and the IEEE TCMC Rising Star Award 2019. He is an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA.

Wengang Zhou (Senior Member, IEEE) received the B.E. degree in electronic information engineering from Wuhan University, China, in 2006, and the Ph.D. degree in electronic engineering and information science from the University of Science and Technology of China (USTC), China, in 2011. From September 2011 to 2013, he worked as a Postdoctoral Researcher with the Department of Computer Science, The University of Texas at San Antonio. He is currently a Professor at the Department of Electronic Engineering and Information Science (EEIS), USTC. His research interests include multimedia information retrieval, computer vision, and computer game.

Houqiang Li (Fellow, IEEE) received the B.S., M.Eng. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively.

He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He has authored and coauthored over 200 papers in journals and conferences. His research interests include multimedia search, image/video analysis, and video coding and communication. He is the Winner of the National Science Funds (NSFC) for Distinguished Young Scientists. He was a recipient of the National Technological Invention Award of China (second class) in 2019 and the National Natural Science Award of China (second class) in 2015. He was also a recipient of the Best Paper Award for VCIP 2012, the Best Paper Award for ICIMCS 2012, and the Best Paper Award for ACM MUM in 2011. He served as the TPC Co-Chair for VCIP 2010, and will serve as the General Co-Chair for ICME 2021. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013. He is a Distinguished Professor of Changjiang Scholars Program of China and the Leading Scientist of Ten Thousand Talent Program of China.

Tao Mei (Fellow, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is currently the Vice President of JD.COM and the Deputy Managing Director of JD AI Research, where he also serves as the Director of the Computer Vision and Multimedia Laboratory. Prior to joining JD.COM in 2018, he was a Senior Research Manager with Microsoft Research Asia, Beijing, China. He has authored or coauthored over 200 publications (with 12 best paper awards) in journals and conferences, 10 book chapters, and edited 5 books. He holds over 25 U.S. and international patents. He is a Fellow of IAPR in 2016. He is the General Co-Chair of IEEE ICME 2019, and the Program Co-Chair of ACM Multimedia 2018, IEEE ICME 2015, and IEEE MMSP 2015. He is or has been an Editorial Board Member of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Pattern Recognition*. He is a Distinguished Scientist of ACM in 2016 and a Distinguished Industry Speaker of the IEEE Signal Processing Society in 2017.