Temporal Meta-Adaptor for Video Object Detection

Chi Wang[1,2]
cwang38@qub.ac.uk
Yang Hua[1]
y.hua@qub.ac.uk
Zheng Lu[2]
steven@anyvision.co
Jian Gao[1,2]
jgao05@qub.ac.uk
Neil Robertson[1]
n.robertson@qub.ac.uk

[1] EEECS/ECIT
Queen's University Belfast
Belfast, UK

[2] Anyvision
Belfast, UK

Abstract

Detecting objects in a video can be difficult due to occlusions and motion blur, where the output features are easily deteriorated. Recent state-of-the-art methods propose to enhance the features of the key frame with reference frames using attention modules. However, the feature enhancement uses the features extracted from a fixed backbone. It is fundamentally hard for a fixed backbone to generate discriminative features for the frames of both low and high quality. To mitigate this challenge, in this paper, we present a meta-learning scheme that learns to adapt the backbone using temporal features. Specifically, we propose to summarise the temporal feature into a fixed size representation, which is then used to make the backbone generate adaptively discriminative features for low and high quality frames. We demonstrate that the proposed approach can be easily incorporated into latest temporal aggregation approaches with almost no impact on the inference speed. Experiments on ImageNet VID dataset show a consistent gain over state-of-the-art methods.

1 Introduction

Video object detection is a more challenging detection task than detection in still images. The reason is that the quality of each single frame can be affected by the capturing process, such as abrupt relative movement between the camera and the target object, or camera out-of-focus. Recent temporal aggregation methods [3, 6, 7, 36, 38, 44] propose to use feature-level temporal information, i.e., features from reference frames in the same video, to enhance the current deteriorated features. Specifically, the attention module [33] is applied to aggregate features from reference frames to the key frame. This significantly improves the performance of video object detection due to the use of global [3, 38] and/or local [7, 44] temporal information.

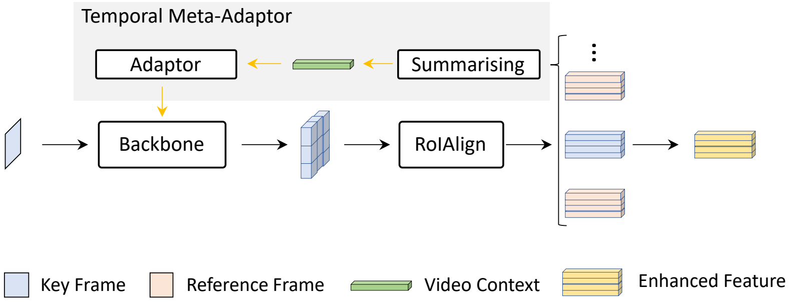

Figure 1: Proposed Temporal Meta-Adaptor integrated with a Faster-RCNN based temporal aggregation model. Some layers are omitted for simplicity.

While being considerably better than a single frame detector, these temporal aggregation methods use a backbone with a fixed network weights to extract features which are then aggregated for feature enhancement. We argue that, however, using a fixed backbone is not ideal due to the vast diversity of video contents, especially when some of the video frames are of significant low quality. Although recovering the lost information from temporal feature is the main purpose of those approaches, the recovering process is heavily dependant on the quality of the fixed backbone output. Temporal aggregation can be hardly reliable if the feature maps extracted are not very informative, either the aggregation is done directly on feature maps [6, 36, 44], or on proposal features [3, 7, 38].

To address this issue, we propose to make the backbone more flexible by conditioning it on temporal features. We present an end-to-end meta-learning scheme named Temporal Meta-Adaptor (TMA) to condition the backbone on temporal features, shown in Fig. 1. TMA comprises two sub-modules. The first module summarises the temporal features into a fixed size representation, which we refer to as video context. Fed by the video context, the second module is learned to adapt the network features for frames of different quality. Specifically, we propose learnable affine transformations working on the internal feature maps to achieve this adaptation. Our TMA can work as a plug-and-play module which can easily be adapted to existing feature memory based temporal aggregation approaches [3, 6, 38]. It results from the fact that the summarising process can smooth out the effect of deteriorated features, which is more robust than directly using them in the attention module. This design allows us to achieve significant performance gain while maintaining the almost same inference speed.

To summarise, we propose three main contributions: (1) To our knowledge, we are the first to build an adaptive backbone to improve the discriminability for frames of different quality via meta-learning in video object detection. (2) We demonstrate that the proposed light-weighted TMA can be easily integrated into the latest state-of-the-art video object detection methods. (3) We validate the proposed methods on ImageNet VID dataset, showing consistent improvements with negligible additional computational cost.

2 Related Work

2.1 Feature Aggregation for Video Object Detection

Recently, feature aggregation methods are proposed to leverage knowledge from reference frames in the same video to enhance the detection performance of the current one. Depending on the range of the aggregation, recent methods can be categorised into three approaches. Firstly, local aggregation [7, 36, 44] uses temporally close information from neighbouring frames. The aggregation can be guided by optical flow or learned relation between detection boxes. Secondly, global aggregation methods [6, 38] consider the semantic relations between detection boxes, which enables a longer range of knowledge transfer than the local methods. Lastly, MEGA [3] combines the local and global aggregation to further enhance the object features.

These approaches mainly focus on aggregating features from reference frames in the same video. The aggregation is achieved by attention modules that discover relationship among temporal features. In this paper, we present an alternative approach to use the temporal features. Instead of using them directly in temporal aggregation, we first condition the backbone on summarised temporal features. The conditioned backbone then produces more discriminative feature for aggregation.

2.2 Video Frames as Context

Context information in video is also studied in topics like video action recognition [20], prediction [19], retrieval [12], etc. To extract the most related context, a number of techniques have been applied such as spatio-temporal convolution [30], RNN [2], non-local/attention [37], correspondence proposals [21], etc. However, these methods mainly work on relations between spatio-temporal features and do not consider video level information. In contrast, our target is to obtain a holistic single video context to guide the backbone.

2.3 Relation to Meta-Learning

The idea of conditioning a neural network using a few observations is also related to meta-learning, where a meta-model is learned to quickly adapt to novel tasks using a few samples [11, 34]. A typical application of meta-learning is few-shot image classification, in which the labelled samples are used as context to build the classifier for corresponding categories. [14, 29]. Garnelo *et al.* [13] propose to condition the prediction model on observations by aggregating them into a fixed size embedding with a symmetric function. This relates to PointNet [25] and DeepSets [41] which also operate on unordered data. As mentioned in [35], there is a connection between methods that using symmetric function [13, 25, 41] and the attention approach [32].

Recent meta-learning approaches further improve the few-shot learning capability by imposing more conditioning. Labelled samples are encoded to guide the feature extractor [23] or directly predict model parameters [28]. Lee *et al.* [18] further apply context-dependent learning rate to have more control over model updating.

The idea of conditioning in meta-learning is related to the proposed method in the sense that learning a fixed single model that works for all scenarios can be overly difficult and may result in worse performance for each one of them, especially when there are conflicts among them [1]. However, the conditioning in meta-learning mainly focuses on utilising

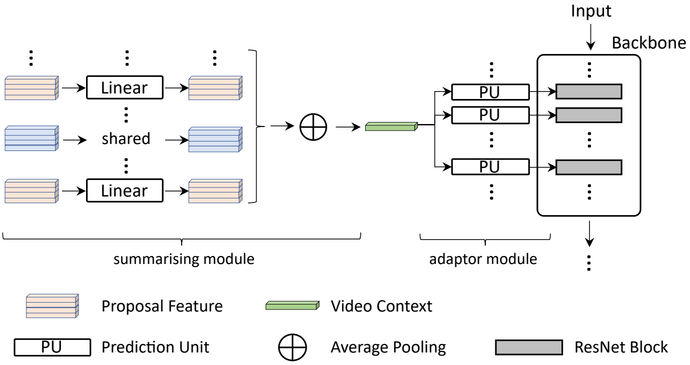

Figure 2: Design of the summarising module and adaptor module. Details can be found in §3.3.

limited labelled data to adapt the model to be more discriminative to similar unlabelled test samples. While in our approach, the conditioning is on extracted proposal features without considering their labels and the purpose is to guide the backbone to be more video-specific.

2.4 Conditional Convolutions and Dynamic Neural Networks

Different from a standard convolutional layer that has fixed weights, a dynamic neural network is conditional on its input, which can be used to increase network capacity [4, 17, 40]. These approaches conditions the network on the current input and normally contains a self-attention module. However, in our approach, the conditioning is on the previously extracted features and the conditioning parameters are directly predicted by a light-weight network.

Another type of conditioning in neural networks is conditional Batch Normalisation [9, 16], where the internal feature is modulated by a set of scaling and shifting parameters. This is commonly used in multi-modality learning [5, 24], style-transfer [9, 16], and few-shot classification [23]. We generally follow the practice of these works on modulating the internal features but with a completely different motivation. To our knowledge, we are the first to modulate the intermediate features in the backbone to improve the discriminability for frames of different quality in video object detection.

In order to prevent domain shift in online learning, Zhang *et al.* [43] use a Kalman filter to smooth the changing of BN statistics. In our case, the shifting and scaling parameters predicted by the video context is much more stable so that it can be easily constructed using a few frames and safely reused afterwards without performance loss.

3 Methodology

Different from a still image which tend to be clearer and better object-centred, objects in a video may be occluded, cropped, motion-blurred or out of focus. For robust object detection in videos, proposal features from reference frames are often used to enhance the feature in the target frame. This is typically achieved by an attention module, where the relation between proposals is used to guide the feature aggregation.

In this section, in complementary to aggregating proposal features, we present an approach which condition the backbone using information from reference frames. We first introduce a light-weight video summarising module to produce a video context. Then an adaptor module adjusts the backbone for the key frame using the obtained video context.

3.1 Summarising Module

We denote a backbone network f_θ to extract features from frames, and our conditioned backbone as f_{θ_c}, where c is a video context, summarised using frames in the same video. The summarising process can be denoted as:

$$c = s(x_{\pi(1)}, x_{\pi(2)}, ..., x_{\pi(M)}), \tag{1}$$

where π is a permutation of the video frames, M is the number of temporal features x_*. The output of summarising function s should be invariant to π, since we consider the video level knowledge. Inspired by DeepSets [41], we construct a summarising function as follows:

$$c = g(h(x_1), h(x_2), ..., h(x_M)), \tag{2}$$

where g is a single symmetric function which is invariant to the input order, such as max pooling and average pooling and h is a linear/non-linear projection function which improves the representation capability of the summarising module.

3.2 Adaptor Module

After obtaining the video context c containing semantic information in the current video, an adaptor module is designed to aid the backbone to enhance the discriminability of the encoded feature using the video context.

Concretely, the adaptation is conducted by applying affine transformations on the features after a selection of convolutional layers. For each of them, a set of video specific parameters $\{\beta, \gamma\}$ (channel-wise shifting and scaling parameter) are predicted using the video context:

$$\beta = PU_0(c), \gamma = PU_1(c), \tag{3}$$

where PU_0 and PU_1 are two dedicated prediction units for β and γ respectively. An input feature a is then modulated to $\hat{a}$ via the following affine transformation using the predicted parameters:

$$\hat{a} = \gamma \circ a + \beta, \tag{4}$$

where $\circ$ and $+$ are channel-wise multiplication and summation.

3.3 Implementation Details

Our method is simple and efficient which can be plugged into any video object detection method with a temporal aggregation module. For experimental purposes, we use SELSA [38] as our base model for its simplicity and competitive performance. The same architecture of the baseline methods, including feature extractor and detection heads, are adopted as per specified in the original paper. We also use the same training hyper-parameters. ResNet-101 and ResNeXt-101 feature extractors are engaged to evaluate the proposed method.

We follow the training procedure of current common practice where temporal aggregation is applied [3, 38]. Two support frames are randomly selected from the same video. Combined with the current target frame, they are used to calculate the video context in the

Methods	Backbone	ms/img	mAP(%)@motion			
			slow	medium	fast	all
SELSA [38]	ResNet-101	65.7	88.93	81.81	63.74	82.85
SELSA + TMA	ResNet-101	67.2	**89.22**	**83.32**	**65.96**	**84.07**
SELSA [38]	ResNeXt-101	104.2	89.69	83.36	65.06	83.63
SELSA + TMA	ResNeXt-101	105.5	**91.32**	**84.70**	**65.43**	**85.02**

Table 1: TMA results on ImageNet VID.

summarising module. The project function h in the summarising module is selected to be the combination of a single linear projection from 1024-d to 1024-d and an average pooling. Comparison with a non-linear projection is discussed in §4.3.1.

Adapting parameter β and γ are predicted using the video context. Two prediction units (PU), for β and γ respectively, are added after each BatchNorm layer in *stage-4* of the ResNet-101 backbone. In each PU, there are two linear layers each followed by a ReLU and a GroupNorm [39] (GN) layer. We use skip connections to link the input of the second linear layer to the output of the second GN layer. Although the PUs can be trained without GN layers, we empirically found that the model with GN achieves better performance. See §4.3.2 for details. The detailed structure of a PU is illustrated in the supplemental material.

TMA is trained jointly with the base detection model. We first turn off gradient tracing to obtain temporal ROI features. Then we turn it back on, feed the ROI features to the summarising module and continue the standard training process. TMA does not require any change in the inference process, as long as there is a memory to save previous features from frames in the same video, which will be used by the summarising module to predict a video context. It is also not necessary to run the video context prediction for every target frame. For example for the ImageNet VID dataset [27], it is sufficient to keep and reuse the calculated context as long as it is calculated using at least 10 different frames in the current video.

4 Experiments

4.1 Dataset and Evaluation Setup

Same as recent approaches [3, 7, 38], a mixture of ImageNet VID and DET [27] dataset is used to train the baseline detection models. ImageNet VID dataset has 3,862 videos in the training split and 555 in the validation split. We aim at learning video-specific knowledge, so the video context is extracted per video and is not shared with others. The support frame is always obtained from the same video as the target frame. Following a common protocol, we use *mean average precision* (mAP) @IOU=0.5 as our evaluation metric. We also evaluate the motion specific mAPs on the validation set. No post-processing or augmentation is conducted for all the evaluations.

4.2 TMA with Temporal Aggregation

We first evaluate the performance gain of adding TMA to SELSA [38], which is one of the state-of-the-art temporal aggregation method for video object detection. The result is shown in Table 1. Under the ResNet-101 backbone, TMA improves 1.2 overall mAP over SELSA with negligible overhead (only 1.5ms slower per image). Especially in the fast motion case,

Method	Backbone	mAP(%)
FGFA[44]	ResNet-101	76.3
MANet[36]	ResNet-101	78.1
D&T†[10]	ResNet-101	79.8
SELSA*[38]	ResNet-101	82.9
RDN[7]	ResNet-101	81.8
MEGA[3]	ResNet-101	82.9
SELSA + TMA	ResNet-101	**84.07**
D&T† [10]	ResNeXt-101	81.6
SELSA*[38]	ResNeXt-101	83.6
RDN[7]	ResNeXt-101	83.2
MEGA[3]	ResNeXt-101	84.1
SELSA + TMA	ResNeXt-101	**85.02**

Table 2: Comparison with state-of-the-arts. *indicates results are from our implementation. †indicates post-processing is used.

Method	mAP(%)
Faster-RCNN [26]	75.80
Faster-RCNN + SE	75.72
Faster-RCNN + SE (shift+scale)	76.40
Faster-RCNN + TMA	**78.29**
SELSA [38]	82.85
SELSA + SE	82.54
SELSA + SE (shift+scale)	82.33
SELSA + TMA (ref. train, self-only test)	83.34
SELSA + TMA (self-only train + test)	83.46
SELSA + TMA	**84.07**

Table 3: Ablation study on the benefit of summarising temporal information as video context.

TMA improves the mAP by 2.2, which shows its superiority for such challenging scenarios. When equipped with a stronger backbone - ResNeXt-101, the overall performance gain using TMA is even higher, improving the mAP by 1.4.

By combining the proposed TMA with the current temporal aggregation methods, we achieve the best result to date on ImageNet VID dataset. The results are shown in Table 2. Please note that no post-processing and data-augmentation are used in our results.

4.3 Ablation Study

In this subsection, we study how TMA helps to improve the video detection performance. We analyse the summarising module and the adaptor module respectively. In all the following studies we use ResNet-101 as backbone.

4.3.1 Summarising Module

Benefit of summarising temporal features. In Table 3, we study the effectiveness of summarising temporal features into a video context from two perspectives: (1), does

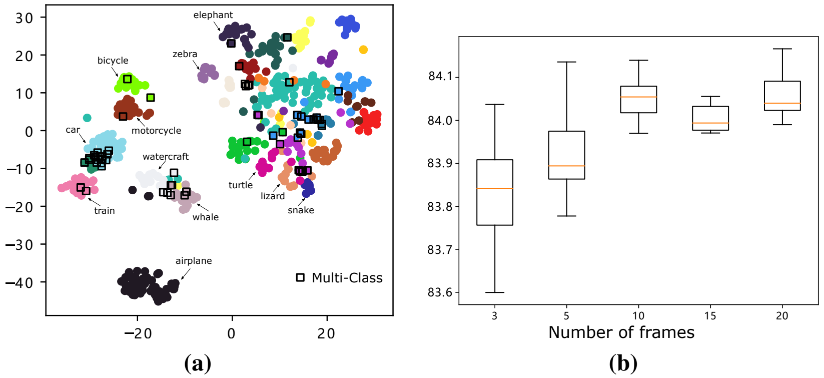

Figure 3: (a) t-SNE visualisation of video context summarised from temporal features. Video context associated with multiple class labels is marked by a black box. (b) Distribution of mAP using different number of frames to summarise video context.

max	average	linear	non-linear	mAP(%)
✓		✓		**84.08**
	✓	✓		84.07
	✓		✓	83.80

Table 4: Ablation study on design choices for summarising module.

it actually improve the discriminability of the backbone? (2), how does it compare with conditioning the backbone only on the current input?

From the first perspective, we apply the proposed conditioning on a single frame detector. The performance gain shows that the conditioning on video context is helpful even without temporal aggregations.

From the second perspective, we compare our full model with two different settings of self-only TMA model: a) training with summarised temporal features and testing with the key frame only, denoted by (ref. train, self-only test), b) training and testing with only the key frame, denoted by (self-only train + test). Both of these settings are worse than the full TMA model that utilises more temporal information.

We also compare with two settings of Squeeze-and-Excitation [15] (SE) block which conditions feature on itself: a) using standard SE Block that only scales the input feature, b) modified SE block that conduct scaling and shifting as our approach. For a fair comparison, in both of the above settings, conditioning is applied in the same layer as our approach.

To verify the effectiveness of adding SE blocks, we first compare with a single-frame detector Faster-RCNN. The first setting achieves no gain but the second setting works slightly better which shows that the added SE blocks do improve the backbone performance. Then we compare using SELSA. The performance is slightly worse in both of the settings, indicating that simply conditioning feature on itself does not work well with temporal aggregation.

Design choices. Table 4 shows the ablation on design choices for aggregating proposal features from multiple video frames. Using single linear projection works slightly better than a more complex non-linear function. For the symmetry function, different choices achieve almost the same performance.

Visualisation of video context. To investigate the information summarised from the

PU structure	mAP(%)
2-Linear	83.81
3-Linear	83.64
2-Linear + GroupNorm	84.07
3-Linear + GroupNorm	84.14

Table 5: Ablation study on difference designs of PU. n-Linear indicates that n fully connected residue blocks are used.

stage	mAP(%)
2	82.90
3	83.53
4	84.07
3, 4	83.91
2, 3, 4	83.43

Table 6: Ablation study on adapting different layers using the TMA in the feature extractor. In each of the stage, features after each BatchNorm layers are adapted.

proposal features. In Fig. 3 (a), we draw t-SNE [31] visualisation on all 555 video context obtained from corresponding videos in ImageNet VID validation set. Ground truth label are used to colour the embeddings. If a video clip contains labels from more than one category, the colour of the dot is chosen by the primary category and we additionally draw a black box around it.

We can observe that man-made categories such as air planes and cars are well separated from natural ones such as turtle and zebra. Correlated categories such as bicycle and motorcycle, whale and watercraft are positioned closer than others. Due to the space limitation, only a few categories are marked. The full figure with all 30 category legends can be found in the supplemental material.

Number of frames for summarising. Because summarised video context vector represents the semantic information of the entire video, it can be reused once it is accurately constructed. As the inference goes, we stop summarising after having processed a certain number of frames. The obtained video context is then reused until the next video begins. In Fig. 3 (b), we show the number of frames needed to obtain an accurate video context. In ImageNet VID dataset, 10 frames are enough for a stable performance.

4.3.2 Adaptor module

Choice of adaptation layer. Layers in the backbone have different receptive fields and thus focus on different levels of details. We add the proposed adaptor module on various layers to adapt a wide range of representations. The results in Table 6 show that TMA has more positive effect when we apply the adaptor module on the last stages of the backbone. This also proves that the semantic information from video context is the key to the performance gain, since the deeper layers of the backbone extract more semantic information [8, 22, 42].

PU Design. A Prediction Unit (PU) predicts the modulation parameter for internal features. As shown in Table 5, although using a simple MLP with two linear layers works very well, adding a GroupNorm [39] layer further boosts the performance. In the same table, we also examine the number of linear layers to find a proper complexity of the PU. The result shows that using two linear layers is sufficient to predict the target parameters we need.

4.4 Complexity Analysis

The computational cost of the proposed module comes from two parts: a) the summarising module which comprises a linear projection and a pooling function. b) the Prediction Unit that predicts adapting parameters using video context. As shown in Fig. 3 (b), the video context can be reused, which means the adapting parameters can also be reused. We reuse the video context by default. As shown in Table 1, the overall computational cost is less than 2ms per frame.

5 Conclusion

In this paper, we propose a meta-learning scheme that learns to adapt the backbone in a video object detection model. We identify that the previous feature aggregation methods are limited by the fixed backbone. When the input frames are deteriorated, a fixed backbone output less discriminable features which affects the temporal aggregation in later stages. We solve the problem by introducing a Temporal Meta-Adaptor which adapts the backbone parameters using a video-level context feature summarised from temporal features. We show that the proposed method can be easily integrated into state-of-the-art methods and achieves considerable performance boosts with almost negligible computational cost.

References

[1] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning. In *CVPR*, 2020.

[2] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *ECCV*, 2018.

[3] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *CVPR*, 2020.

[4] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020.

[5] Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *NIPS*, 2017.

[6] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *ICCV*, 2019.

[7] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *ICCV*, 2019.

[8] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[9] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017.

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[12] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020.

[13] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *ICML*, 2018.

[14] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NIPS*, 2016.

[18] Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *ICLR*, 2020.

[19] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *CVPR*, 2021.

[20] Xianhang Li, Yali Wang, Zhipeng Zhou, and Yu Qiao. Smallbignet: Integrating core and contextual views for video classification. In *CVPR*, 2020.

[21] Xingyu Liu, Joon-Young Lee, and Hailin Jin. Learning video representations from correspondence proposals. In *CVPR*, 2019.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[23] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2017.

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[28] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[30] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[32] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NIPS*, 2017.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[34] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016.

[35] Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *ICML*, 2019.

[36] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *ECCV*, 2018.

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[38] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *ICCV*, 2019.

[39] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

[40] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *NeurIPS*, 2019.

[41] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *NIPS*, 2017.

[42] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*. Springer, 2014.

[43] Zhenyu Zhang, Stéphane Lathuilière, Elisa Ricci, Nicu Sebe, Yan Yan, and Jian Yang. Online depth learning against forgetting in monocular videos. In *CVPR*, 2020.

[44] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017.