

Improving Video Instance Segmentation via Temporal Pyramid Routing

Xiangtai Li, Hao He, Yibo Yang, Henghui Ding, Kuiyuan Yang, Guangliang Cheng,
Yunhai Tong✉, Dacheng Tao, *Fellow, IEEE*

Abstract—Video Instance Segmentation (VIS) is a new and inherently multi-task problem, which aims to detect, segment, and track each instance in a video sequence. Existing approaches are mainly based on single-frame features or single-scale features of multiple frames, where either temporal information or multi-scale information is ignored. To incorporate both temporal and scale information, we propose a Temporal Pyramid Routing (TPR) strategy to conditionally align and conduct pixel-level aggregation from a feature pyramid pair of two adjacent frames. Specifically, TPR contains two novel components, including Dynamic Aligned Cell Routing (DACR) and Cross Pyramid Routing (CPR), where DACR is designed for aligning and gating pyramid features across temporal dimension, while CPR transfers temporally aggregated features across scale dimension. Moreover, our approach is a light-weight and plug-and-play module and can be easily applied to existing instance segmentation methods. Extensive experiments on three datasets including YouTube-VIS (2019, 2021) and Cityscapes-VPS demonstrate the effectiveness and efficiency of the proposed approach on several state-of-the-art instance and panoptic segmentation methods. Codes will be publicly available at <https://github.com/lxtGH/TemporalPyramidRouting>.

Index Terms—Dynamic Network, Temporal Modeling, Video Instance Segmentation, Video Scene Understanding

1 INTRODUCTION

MODELING dynamic video inputs for scene understanding is a fundamental vision research direction [1], [2]. Video Instance Segmentation (VIS) [1], has been proposed and drawn lots of attention. VIS aims to simultaneously classify, segment and track object instances in a given video sequence, which is a very challenging task that involves the instance correlations along the temporal dimension.

Previously, many state-of-the-art methods [3], [4] model the temporal correlation by feature aggregation. For example, Mask-Prop [3] proposes to warp mask features via a strong object detector [5] while VisTR [4] treats the entire video clips as sequence inputs and aggregates features via the transformer detector [6]. Moreover, bottom-up method [7] adopts separate object instances by clustering learned pixel embeddings [8] then performs tracking and association via specific post-processing. However, these methods pay little attention to the scale variation problems when modeling temporal feature aggregation, resulting in ambiguous predictions for large variation objects. As shown in Fig. 1(a), the cow in one video clip has a large-scale variation between frames. To segment and track objects consistently across consecutive frames in a video, it is essential to incorporate past context and feature pyramids through a temporal model. The scale variation problems have been well studied in the image domain via pyramid representation, as shown in Fig. 1(b). Feature Pyramid

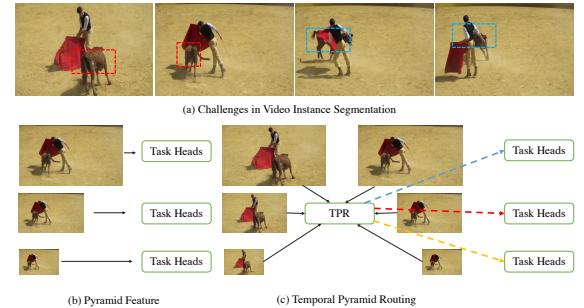


Fig. 1: (a) The illustration of problems in video instance segmentation: scale variation (red boxes) and overlapping (blue boxes). (b) Pyramid feature learning in image understand tasks. (c) Our proposed Temporal Pyramid Routing. Each line in a different color represents per pixel-wise routing across different scales along the temporal dimension.

Network (FPN) [9], [10], [11] and its variants [12], [13], [14] are the most classic architectures to establish a pyramid network for object representation. It assigns instances to different pyramid levels according to the object sizes. However, it has *not* been well explored in the video on how to align other feature pyramids among the different feature levels along the temporal dimension. Also, it has not been well studied on how to model the spatial-temporal feature pyramids for the down-stream tasks.

Several works utilize optical flow [15], [16] or non-local operators [17] to aggregate temporal features for video object detection. However, both limited scale variation and simple scenario on ImageNet VID dataset [18] make current methods mainly focus on single scale feature aggregation using Faster R-CNN [19] as the baseline. In this work, we aim to design an approach that enhances the network's ability to handle the video domain's scale

- X. Li and Y. Tong are with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. This work is supported by the National Key Research and Development Program of China (No.2020YFB2103402).
- H. He is with the Department of National Laboratory of Pattern Recognition, Institute of Automation, Beijing, China.
- Y. Yang and D. Tao are with JD Explore Academy, Beijing, China.
- G. Cheng is with SenseTime Research, Beijing, China.
- H. Ding is with ETH Zurich, Switzerland.
- K. Yang is with Xiaomi, Beijing, China.

variation problem.

Directly fusing the previous feature pyramid into the current corresponding feature pyramid is a trivial solution because there may be significant motion variation and background noise in the video. Thus, filtering irrelevant features out and dynamically assigning relevant parts in the previous frame is essential for temporal consistent feature representation. However, existing warping-based approaches [16], [17], [20] may lead to high computation costs. Our method is inspired by the success of dynamic routing in image tasks [21], [22]. It performs the pixel-wised fusion across the feature pyramid, controlled by adaptive learned gates. Compared with static network, it has the advantage of modeling various inputs since the network architecture can be changed according to the inputs. We extend such dynamic network into the video scene understanding and propose a conceptually novel method for temporal feature learning, called Temporal Pyramid Routing (TPR). It dynamically aligns and fuses features, where the pixel-wise routing is performed between adjacent feature pyramids. As shown in Fig. 1(c), our TPR takes two adjacent temporal feature pyramids as input and outputs refined feature pyramids for the down-stream task.

In particular, we design a novel pyramid routing space that mainly contains two steps: Dynamic Aligned Cell Routing (DACR) and Cross Pyramid Routing (CPR). The former uses specifically designed double gates to filter out background noise and absorb relevant semantics. The latter introduces the aligned features from the previous frame into the remaining scales of the current frame dynamically and efficiently. Fig. 2 gives a visual example of the TPR inference procedure. For each pyramid feature, only highlight parts from the previous frame are propagated into the current frame via pixel-level routing, which avoids the complex background noise and redundant information. Finally, we verify our proposed TPR on various baselines [23], [24], [25] on the YouTube-VIS dataset [1]. The experimental results show that our proposed TPR improves baselines by a significant margin (about 2%-3% mAP) with little extra GFlops. Overall, the proposed TPR is fundamentally different from the existing feature aggregation methods in video. TPR has utilized pixel-wise gated routing while maintaining inference efficiency, while previous works [16], [17] introduce much more computation and redundant information between the frames. With Swin Transformer [26] as backbone, our method can achieve 45.9 mAP (online inference) which outperforms MaskProp [3] with *faster* inference speed (about 10 time faster than MaskProp). Our main contributions are summarized as follows:

- 1) We propose a novel framework named Temporal Pyramid Routing for temporal dynamic multiscale representation learning. In particular, we design two main components, including Dynamic Aligned Cell Routing and Cross Pyramid Routing to aggregate features from the previous frame while keeping efficiency.
- 2) Our TPR is a plug-in module where we append it into several image segmentation methods to perform temporal dynamic routing for two different video segmentation tasks, including video instance segmentation and video panoptic segmentation.
- 3) Detailed experiments and analyses on the YouTube-VIS dataset indicate our proposed TPR's efficacy in improving the accuracy and keeping light-weight. Our method improves baselines by a significant margin. Our method achieves the state-of-the-art results *under the online inference setting*. We further verify the generality of TPR on Video Panoptic Segmentation (VPS).

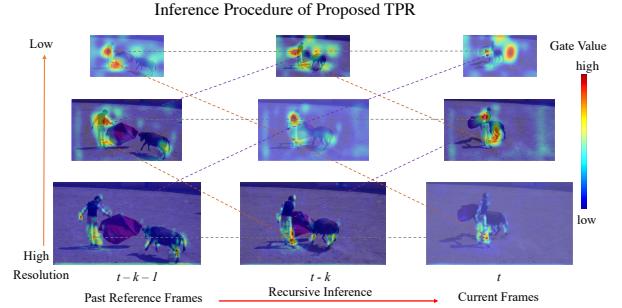


Fig. 2: An illustration of the inference procedure of our proposed TPR. **Only parts of the feature** (highlighted by gates) are propagated into the next frame, *sparsely* and *recursively*.

2 RELATED WORK

Instance Segmentation Instance Segmentation aims to detect and segment each instance [27], [28]. The two-stage pipeline Mask-RCNN and its variants [5], [11], [29] first generate object proposals using Region Proposal Network (RPN) [19] and then predict boxes and masks on each RoI feature. Recently, several single-stage methods [23], [25], [30] achieve significant progress and comparable results with two-stage pipelines. Meanwhile, there are several bottom-up approaches [31]. Our method is built on those top-down works and explores the correlation of the temporal feature pyramid.

Video Instance Segmentation The VIS requires classifying, segmenting each instance for each frame, and assigning the same instance with the same id. MaskTrack-RCNN [1] is the first attempt to address this task, where they also propose YouTube-VIS dataset for benchmarking video instance segmentation algorithms. Maskprop [3] is built on state-of-the-art detection method HTC [5] and crops the extracted features via predicted masks, then propagates them temporally to improve the segmentation and tracking. Based on [1], the work [32] proposes a modified variational autoencoder, while CompFeat [33] aggregates both frame-level and object-level information. Recently, VisTR [4] solves VIS problem using detection transformer [6]. VIS is also closely related to Video Semantic Segmentation [15] and Video Panoptic Segmentation [2]. Multi-Object Tracking and Segmentation (MOTS) task [34] is proposed to evaluate MOT along with instance segmentation. Recently, several works [35], [36], [37] use object query to unify video scene understanding tasks. Due to its limited scale distribution and much fewer object categories, we do not compare it in this paper. TPN [38] also proposes to use multiscale temporal pyramid networks for action recognition. However, this method cannot generalize to video dense prediction tasks due to the misaligned temporal representation.

Dynamic Network Design Several works design the specific dynamic components [39], [40], [41] and architectures [42] for image classification task where the parameters and architectures are conditioned on inputs. Recently, inspired by Neural Architecture Search (NAS) [43], [44], [45], [46], several works propose dynamic network design on the dense pixel prediction tasks such as semantic segmentation [22] and object detection [21]. These methods have data-dependent routes which are adapted from the scale distribution of each input. However, the dynamic network for video dense prediction tasks including VIS and VPS are *not well explored*. Our methods are inspired by those works, but differ from them. In particular, we design several new routing cells to align

and fuse temporal information. Moreover, a new routing space is introduced to perform cross fusion between feature pyramids. To the best of our knowledge, we are the first to design a novel temporal routing strategy for the video segmentation tasks. Moreover, our methods are *fundamentally different* from previous attention based approaches [47] where the computation cost is fixed during inference. Our dynamic architecture propagates parts of features according to the input, which has been shown effective and robust.

3 METHOD

3.1 Overview of Dynamic Routing in Image Domain

Basic Notation and Concepts. Compared with static networks, dynamic networks [21], [22], [40] have larger network capacity and better results under certain budget constraints. In a dynamic network, a *routing space* is defined as *all potential paths*, and the routing process is to aggregate potential multiscale features and choose routing paths conditioned on each input image using *cells* and *gates* in each path. *Cells* are the basic computation blocks, while the *gates* control how much computation is performed conditionally for each input. Formally, given the input feature $X_s^l \in \mathbb{R}^{C \times H \times W}$ with $H \times W$ pixel-level locations and C channels, a cell processes the feature with an operation set \mathcal{O} contains widely-used network operators (e.g., convolutions, identity mapping, pooling),

$$H_s^l = \sum_{O^i \in \mathcal{O}} O^i(X_s^l), \quad (1)$$

where l means the current depth of routing space, $s \in \{2^i, i = 3, 4, \dots, k\}$ denotes the feature scale, and k is the largest down-sample ratio, for example, in FCOS [48], $k=7$, namely P3, P4, ..., P7. The output H_s^l is transferred to different scales according to the gating factor m_s^l . Take pixel-wise routing gates as an example, each pixel-level location has a gate factor generated by

$$m_s^l = G_s^l(X_s^l; \theta_s^l), \quad (2)$$

where θ_s^l is the network parameters learned for defining the gating function $G_s^l(\cdot)$. The output m_s^l is normalized via a variant of the soft differentiable tanh function [21]. Finally, X_s^l is transformed into Y_s^l after going through a path formed by cells and gates, i.e.,

$$Y_s^l = H_s^l \cdot m_s^l. \quad (3)$$

In summary, Dynamic Routing controls the operators applied on an image via gates conditioned on the image content, which is different from NAS-based approaches [43], [46], where the network structure is only varied during training phase but fixed for inference.

Motivation for Video Instance Segmentation. As shown in Fig. 1(a), a video clip contains a sequence of images stacked along the temporal dimension. Compared with a static image, a video contains more but redundant information, and also more variations caused by motion. Conditioned on a video input, dynamic routing is a natural mechanism to select useful instance information for VIS.

3.2 Temporal Pyramid Routing

Temporal Pyramid Dynamic Routing Space. The basic task for dynamic network is routing space design, which determines the potential capacity of the network. In this work, we use feature

pyramid to represent multiscale information of each frame, and feature pyramids from a video to capture temporal information. The routing space is designed to control the information flow along both the temporal and scale dimensions. Feature pyramids go through the routing space, and are conditionally refined by selected paths. The whole routing process is defined as Temporal Pyramid Routing (TPR). Finally, task-specific heads are attached to the refined feature pyramid for down-stream tasks, such as detection, segmentation and tracking. To be noted, it is a plug-and-play module that can be applied to any method using feature pyramid as input, which is verified in Sec. 4.2.

As shown in Fig. 3(a), we take two feature pyramid inputs in one video sequence as an example. One feature pyramid X_s^q is from the query frame q (i.e., current frame, denoted as blue rectangles) and the other feature pyramid X_s^r is from the reference frame r (i.e., support frame, denoted as green rectangles). The two input feature pyramids go through TPR, which routes temporal information via Dynamic Aligned Cell Routing (DPCR) and scale information via Cross Pyramid Routing (CPR). DPCR absorbs the most relevant information from the reference frame in a pixel-wised manner scale by scale and CPR routes the absorbed information to other scales. We will specify the designing details of these two components.

Dynamic Aligned Cell Routing. As mentioned above, it is crucial to design a temporal pixel-wise routing cell to avoid noises and sample the most important parts. We apply a sample and filter strategy to align temporal feature frames. First, Dynamic Aligned Cell Routing utilizes the dynamic sampling operator [49] to sample corresponding pixel locations, and then irrelevant parts are filtered out via our double gates design. Two features $X_s^q(i)$ and $X_s^r(i)$ are the inputs of our Dynamic Aligned Cell, where i is the index of each spatial position. The first step is to find the correspondence between the two features. We concatenate two features and generate two individual offset fields O_s^r, O_s^q via one 3×3 convolution layer. Then the corresponding features are sampled according to the predicted offset fields through deformable convolution (DCN) [49]. Take the query feature $X_s^q(i)$ with predicted offset O_s^q as an example,

$$\hat{X}_s^q(i) = \sum_{p_n \in R} X_s^q(i + p_n + O_s^q(p_n)) \cdot W(p_n), \quad (4)$$

where p_n enumerates the locations in neighbor R and W is kernel weight of DCN. Then we perform pixel-wise dynamic routing using refined \hat{X}_s^q, \hat{X}_s^r and original X_s^q via our double gates.

There are two types of gate in DACR, i.e., inner gate and outer gate. The inner gate aims to control the sampled reference features and find fine-grained supports from reference features, while the outer gate emphasizes informative locations in query frame and only absorbs features from reference frames accordingly. In particular, for inner gates, we choose stride convolution and max-pooling for gate map generation to highlight the most salient areas. Following Equ. 3, the output of inner cell routing is formally defined as,

$$\hat{Y}_s = \{Y_s^i | Y_s^i = H_s^i \cdot \text{Gate}(\text{Cat}(\hat{X}_s^q, \hat{X}_s^r)), i \in q, r\}, \quad (5)$$

where *Cat* means concatenation operation, *Gate* is the inner function and the hidden states H_s^i are generated by cell via Equ. 1 where the inputs are the aligned outputs from Equ. 4. During inference, only pooled areas that are selected by the inner gate can be propagated into \hat{Y}_s^q , which can be efficiently implemented via

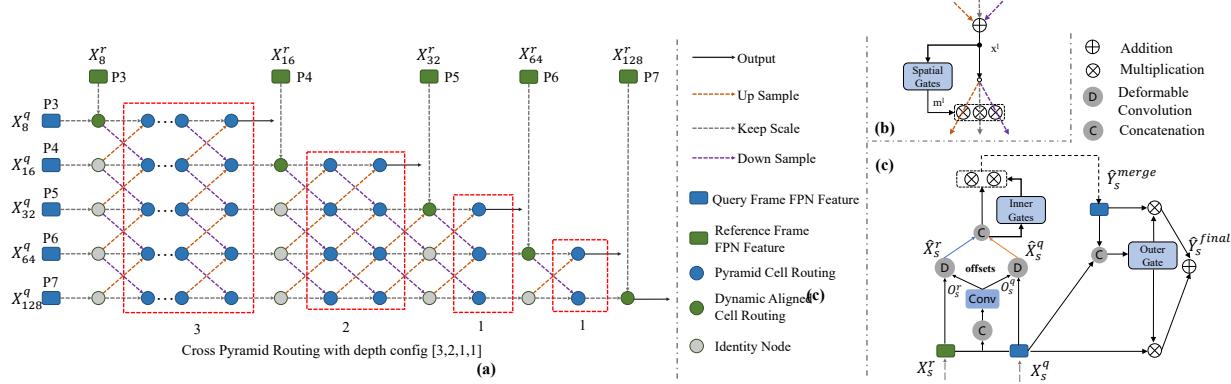


Fig. 3: An illustration of our proposed Temporal Pyramid Routing system, which takes two feature pyramids as inputs (blue rectangles and green rectangles), and outputs the refined feature pyramids for down-stream tasks. Following VIS paper [1], Query frame represents the current frame while Reference frame is randomly sampled from the same video clip around the current frame. (a) Overview of our TPR. It mainly contains two components: Dynamic Aligned Cell Routing (Green Circles) and Cross Pyramid Routing (Blue Circles in Red dotted boxes). (b) Detailed cell operation in Cross Pyramid Routing. (c) Detailed cell operation in Dynamic Aligned Cell Routing.

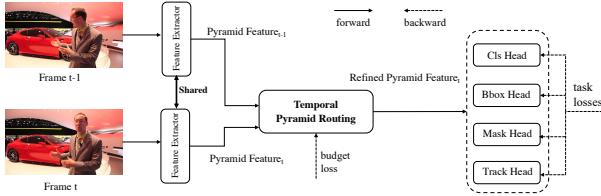


Fig. 4: An illustration of network architecture with our proposed TPR. The feature extractor contains backbone and FPN. During inference, Frame $t - 1$ is used as reference frame. Best view it on screen.

masked convolution [50]. Finally, we fuse two features Y_s^q and Y_s^{merge} via 1×1 convolution, and denote the fusion result as \hat{Y}_s^{final} for short.

Although the inner gates can be effective to avoid background noises, the contents are flexibly dependent on sampled offsets O_s^r, O_s^q , which affect the stability of the query frame feature learning. Thus, we propose to combine query frame feature X_s^q and merged feature Y_s^{merge} via an outer gate map G_s . The gate map G_s is generated by convolutional fusion of X_s^q and Y_s^{merge} followed by a sigmoid function. The convolutional fusion is simply implemented by two 1×1 convolution. Then, the final feature map is obtained as the weighted sum of X_s^q and Y_s^{merge} according G_s ,

$$\hat{Y}_s^{final} = G_s \cdot X_s^q + (1 - G_s) \cdot Y_s^{merge} \quad (6)$$

Cross Pyramid Routing. After harvesting reference features from each scale s , the next important problem is how to distribute \hat{Y}_s^{final} to other scales of query feature pyramid. In FPN-like networks [9], [12], semantically strong features in low-resolution and semantically low features in high-resolution are combined in the network, which are also used in most state-of-the-art image instance segmentation approaches [23], [25]. CPR follows the cross-scale, pixel-wise routing strategy [21] with two exceptions. One is the input feature that is modified by DACR. The other is the lower-triangular shaped routing space, where lower-level features go through deeper paths to upper-level features, to bridge the semantic gap. The default routing depth l is [3, 2, 1, 1] is used in all experiments, more choices about l can be found in

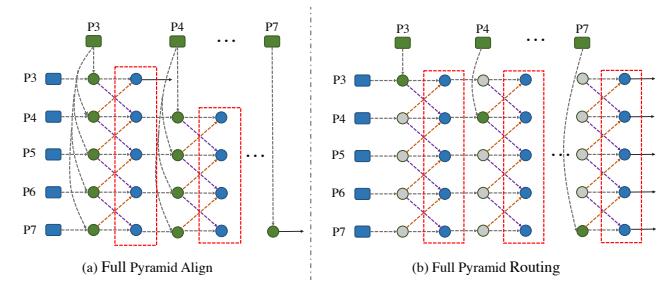


Fig. 5: An illustration of more routing space designs. (a). Full pyramid align. (b). Full pyramid routing.

the supplementary. The detailed procedure of each routing cell is shown in Fig. 3(b). Both DACR and CPR work at the same time for both training and inference.

In addition to the proposed CPR design, there are another two straightforward design choices. One is Full Pyramid Align routing as illustrated in Fig. 5(a), where each node from the reference feature pyramid is aligned to all nodes in the query feature pyramid via DACR. The other is Fully Pyramid Routing as proposed in [22], where all nodes in query feature pyramid are with the same routing depth. The former suffers from the severe misalignments between features across different scales. The latter treats all scales equally and leads to more redundant computation for upsampling high-level features. Moreover, the direction of CPR is **not** top-down. There are three reasons: Firstly, adopting bottom-up routing for CPR has better results due to the complementary property of FPN (top-down). Secondly, it is also beneficial for aligned high resolution representation propagation, since misalignment in low resolution is severer [51]. Thirdly, bottom-up routing also achieves faster inference speed since high resolution features go out in the earlier stages. More detailed results can be found in Sec. 4.2.

Discussion With Warping Based Approaches. Compared with previous works using optical flow for warping features [15], [16], our dynamic network does not require the extra optical flow network training and adaption. Compared with DCN sampling [52], our proposed DACR can avoid background noises and achieve faster inference speed since only sparse locations are involved during the inference.

3.3 Network Architectures

Extending Image Instance Segmentation into Video VIS requires the association of instances in a video. Following previous works [1], [24], we also add a tracking head for fair comparison, as shown in Fig. 4. The tracking head outputs feature embedding of each instance candidate. Suppose there are K instances identified from previous frames, the candidate box i in the current frame will be assigned to a label k according to the assignment probability,

$$p_i(k) = \begin{cases} \frac{e^{f_i^T f_k}}{1 + \sum_{j=1}^K e^{f_i^T f_j}} & k \in [1, K], \\ \frac{1}{1 + \sum_{j=1}^K e^{f_i^T f_j}} & k = 0, \end{cases} \quad (7)$$

where $1 \leq k \leq K$ indicates the candidate is associated to one of the K instances and $k = 0$ means the candidate is treated as a new track, f_i and f_j ($j \in [1, K]$) denotes the feature embedding of the candidate and the K pre-identified instances, respectively. Then we perform the cross-entropy loss by treating it as a multi-class classification problem. As shown in Fig. 4, we insert our proposed TPR between feature extractor and these task-specific heads (detection, segmentation, tracking) for VIS.

Loss Functions Since the dynamic network has a large capacity while we only have limited computational resources, we constrain the computational cost following previous works [21], [22]. We consider all the locations involved in the receptive field of locations with positive gating factors, which are obtained by a max-pooling layer. We denote such computation budget from that gate as B^l for the layer l . The budget loss is normalized by the overall computational complexity C^l .

$$\mathcal{L}_{\text{budget}} = \frac{\sum_l B^l}{\sum_l C^l} \quad (8)$$

Finally, the total loss \mathcal{L} is the combination of tasks-specific heads loss and budget loss. The positive hyper-parameters λ_1 and λ_2 are set to achieve a trade-off between efficiency and effectiveness, where we set $\lambda_1 = 1$ and $\lambda_2 = 1.5$ following [21] by default. The $\mathcal{L}_{\text{tasks}}$ contains detection, segmentation and tracking loss. In our experiments, we find that increasing $\lambda_1 = 1$ hurt the performance since most features from the previous frame are not aligned into current frame during inference.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{tasks}} + \lambda_2 \mathcal{L}_{\text{budget}}. \quad (9)$$

Training and Inference for Video Segmentation Tasks For VIS, following [1], [24], we train our network in a fair setting by randomly sampling one reference frame around the query frame. During inference, we adopt the online setting where only one previous reference frame is considered. Since our TPR can align the feature pyramid across temporal dimension, the reference frame also contains earlier cues in a recursive manner. We adopt the same strategies for tracking heads as MaskTrack R-CNN [1] by maintaining a memory to store the feature vectors of existing instances. For VPS, we follow the original VPS setting where we adopt UPSNet [53] with extra tracking heads [1] as our image baseline and insert our TPR into the neck part of the model.

4 EXPERIMENT

4.1 Dataset and Evaluation Metric

Video Instance Segmentation We mainly evaluate our TPR on YouTube-VIS-2019 dataset [1], which has 2238 training, 302

validation, and 343 test videos. Each video is annotated with per-pixel segmentation, category, and instance ids. The dataset contains 40 object categories. Since only the validation set is available for evaluation, all results reported in this paper are evaluated on the validation set by uploading results to the online server. Following previous work, we adopt mAP proposed in [1] which is average precision (AP) based on a spatial-temporal Intersection-over-Union (IoU) as metrics. Both metrics adopt a COCO format where the AP is computed by averaging over multiple IoU thresholds (from 0.5 to 0.95). This metric considers all three tasks, including detection, segmentation, and tracking in a unified way. Since the dynamic network has different GFlops for different inputs, we report the average inference time (Frames Per Second, FPS) of all validation videos using one V-100 GPU card following [4]. Following [21], we also report different GFlops calculated on the validation set. We also perform experiments on YouTube-VIS-2021 datasets which is an improved and augmented version of YouTube-VIS-2019 dataset. It has 2985 training videos, 421 validation videos and 453 test videos. Following [57], we report our results on validation set.

Video Panoptic Segmentation We adopt the Cityscape-VPS dataset which has 400 training videos and 100 validation videos. Each video consists of 30 consecutive frames, with every 5 frames paired with the ground truth annotations. For each video, all 30 frames are predicted, and only the 6 frames with ground truth are evaluated. We adopt the Video Panoptic Quality (VPQ) as the default metric.

4.2 Experiments on YouTube-VIS for VIS

Overview and Baseline We will first perform ablation studies using BlendMask [23] with ResNet50 [58] backbone (with tracking head in Sec. 3.3) as the baseline to prove the effectiveness of our TPR. Then we will give a detailed analysis and further comparison of our TPR. In addition, we also report performance gain on several other instance segmentation methods [1], [24], [25]. Finally, we compare our methods with previous works at last.

Experiment Settings We use the PyTorch library [59] and Detectron2 [60] to implement all the models. Following [1], [3], all the image instance segmentation models are pre-trained on COCO datasets [61]. We resize the original frame size to 640×360 for both training and testing. We train all the models for 12 epochs. During training, the initial learning rate is set to 0.005 and decays with a factor of 10 after epoch 8 and 11. Multiscale training is adopted to obtain a strong baseline. For each input frame, we randomly select two frames from the same video, one used as the query frame and the other used as the reference frame for training TPR and the tracking head. The reference frame is randomly sampled in a limited range of 3 for the query frame during training. All the ablation results are reported by averaging three runs for YouTube-VIS-2019.

Ablation study for effectiveness of TPR We first verify the effectiveness of each component in TPR in Tab. 1(a). Starting from the baseline model, we obtain 33.8% mAP. Adding DACR results in about 1.4 % mAP gain, while CPR leads to 0.7 % mAP, which indicates temporal cues are more important. After combining both, there is a significant gain over the baseline by 2.1 % mAP. After using a strong baseline with ResNet101 as backbone, our TPR also has 2.5% mAP gain, which proves the scalability of our approach.

Ablation study for DACR In Tab. 1(b), we investigate the necessity of our essential modules for DACR design in Sec. 3.2. We

TABLE 1: Ablation studies. We first verify the effect of each module and design choices in the first row. Then we perform several comparison and analysis of our module in the second row. Due to the limited space, more implementation details can be in supplementary.

(a) Effect of two main components on various baselines.

Backbone	+ DACR	+ CPR	mAP(%)
ResNet50	-	-	33.8
ResNet50	✓	-	35.2
ResNet50	-	✓	34.5
ResNet50	✓	✓	36.2
ResNet101	-	-	36.6
ResNet101	✓	✓	39.1

(d) Comparison other feature warp operations.(w: with)

Sampling Method	mAP(%)
baseline + CPR	34.5
w DACR	36.2
w DCNv2 fusion [54]	35.2
w STSN [52]	35.1
w CBAM [47]	34.8
w optical flow warp [15], [55]	35.0
w feature flow warp [56]	34.4

(b) Effect of each component in DACR.

Method	Equ. 4	Equ. 5	Equ. 6	mAP(%)
Baseline		✓		33.8
			✓	34.5
		✓		34.3
	✓	✓		34.8
	✓	✓	✓	35.2

(c) Ablation on routing space design under the same routing depth config. GFlops are measured on the entire validation set.

Settings	mAP(%)	GFlopsavg
baseline+DACR	35.2	58.9
+CPR	36.2	+2.3
+Full pyramid routing	35.0	+6.5
+Full pyramid align	33.4	+5.9
Top-Down routing(in CPR)	34.6	+4.8

(e) Generalization on other instance segmentation methods.

Settings	mAP(%)
SipMask [24]	33.8
+Our TPR	36.2
YOLACT [25]	32.3
+Our TPR	34.8
MaskTrack-RCNN [1]	30.2
+Our TPR	31.6

(f) Computation Analysis on various baselines. Both GFlops and parameters are reported. GFlops are measured on the entire validation set.

Method	Backbone	mAP(%)	GFlopsavg	GFlopsmax	GFlopsmin	Params
Baseline	ResNet50	33.8	56.8	56.8	56.8	41.2M
	ResNet50	36.2	+4.4	+4.9	+4.2	+10.2M
Baseline	ResNet101	36.6	75.7	75.7	75.7	60.1M
+TPR	ResNet101	39.1	+4.9	+6.1	+4.5	+14.2M

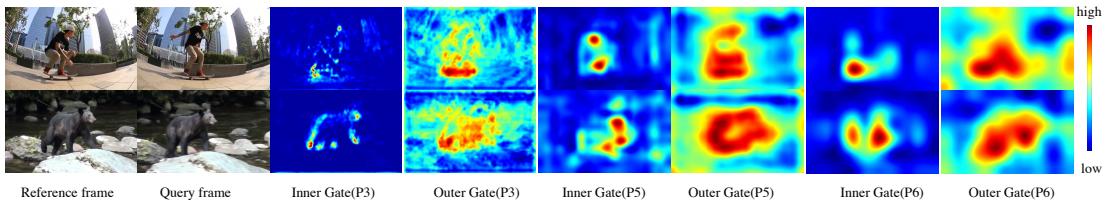


Fig. 6: **Visualization of gates in Dynamic Aligned Cell Routing.** We choose three features (P3, P5, P6). The inner gates are all from the reference frame to control how much information is needed from the previous frame. The outer gates highlight the important regions in the current frame. Best view it on screen and zoom in.

progressively improve the baseline with: (1) DCN-based sampling in Equ. 4; (2) dynamic routing with inner gates in Equ. 5; (3) Outer gates fusion in Equ. 6. Consistent performance improvements can be observed after introducing the above modifications, and we can draw two essential conclusions: (1) Both DCN-based sampled and dynamic routing can align features and have the same impact. However, combining both leads to better results. (2) Outer gates work complementary with the previous two and can further improve the performance.

Ablation study for Cross Pyramid Routing Design We give more routing space design comparison with our proposed Cross Pyramid Routing in Tab. 1(c). We conduct two different types of routing space design introduced in Sec. 3.2 and Fig. 5. Top-Down routing means we carry out routing from high-level features to low-level features and reverse the routing order of CPR. It results in inferior results since misalignment in low resolution is severer [51] and propagating such misaligned features into different scales hurts the performance for down-stream tasks. It also leads to more GFlops. The results indicate the effectiveness and efficiency of our CPR design and routing space design because all other designs will cause more misalignment across different scales and increase the ratio of misalignment parts for each feature pyramid while having more computation cost. The results verify our discussion in Sec. 3.2.

Comparison with Warping-Based approaches In Tab. 1(d), we give a further comparison with several feature warping-based approaches, including using optical flow like warping [16] and DCN-like warping [52]. We use the baseline method with our CPR for fair comparison. Compared with those works, our approach has

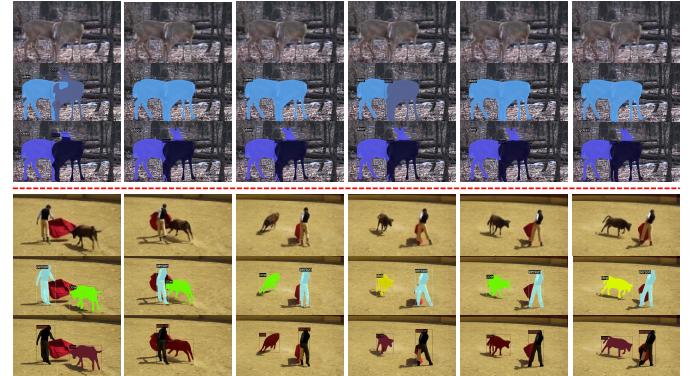


Fig. 7: **Visualization results using our TPR on YouTube-VIS validation set.** Each row has five sampled frames from a video sequence. The first row for each video shows the original frames. The second row illustrates the mask predictions of Baseline Method (BlendMask with Tracking Head) and the third row those obtained with our TPR. Compared to baseline, our TPR tracks object instances more robustly even when they overlap with each other. Note that the same color represents the same object (id). Best view on the screen and zoom in.

the best result, mainly because our proposed double gates can well propagate more relevant information.

Generalization to More Instance Segmentation Approaches We further generate our approaches on more methods, including SipMask [24], MaskTrack-RCNN [11] and YOLACT [25]. All the methods use ResNet50 as backbone. As shown in Tab. 1(e), The

TABLE 2: **Video instance segmentation AP (%)** on the YouTube-VIS-2019 [1] validation dataset. The compared methods are listed by publication date. Note that, for the first three methods, we use the results reported by the re-implementations in [1] for VIS. For BlendMask [23] and SipMask baselines [24], we report our re-implementation results for fair comparison. \ddagger means using DCN [49]. The methods with underline are offline inference, while the others are online.

Method	backbone	FPS	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
DeepSORT [62]	ResNet50	-	26.1	42.9	26.1	27.8	31.3
FEELVOS [63]	ResNet50	-	26.9	42.0	29.7	29.9	33.4
OSMNN [64]	ResNet50	-	27.5	45.1	29.1	28.6	33.1
MaskTrack R-CNN [1]	ResNet50	20.0	30.3	51.1	32.6	31.0	35.5
MaskProp [3]	ResNet50 \ddagger	< 2	40.0	-	42.9	-	-
MaskProp [3]	ResNet101 \ddagger	< 1	42.5	-	45.6	-	-
STEm-Seg [7]	ResNet50	-	30.6	50.7	33.5	31.6	37.1
STEm-Seg [7]	ResNet101	2.1	34.6	55.8	37.9	34.4	41.6
CompFeat [33]	ResNet50	-	35.3	56.0	38.6	33.1	40.3
VisTR [4]	ResNet50	30.0	36.2	59.8	36.9	37.2	42.4
VisTR [4]	ResNet101	27.7	40.1	64.0	45.1	38.3	44.9
BlendMask-baseline	ResNet50	19.3	33.4	52.8	36.1	32.7	38.0
TPR-BlendMask	ResNet50	16.1	36.2	57.2	39.1	36.2	42.3
SipMask-baseline	ResNet50	21.2	33.8	53.1	37.4	35.2	41.0
TPR-SipMask	ResNet50	15.3	36.0	57.0	36.8	36.2	43.6
TPR-BlendMask	ResNet101	10.1	39.1	59.1	43.4	38.8	45.6
TPR-BlendMask	ResNet101 \ddagger	8.2	42.2	63.3	45.7	40.8	49.0
BlendMask-baseline	Swin-tiny	12.3	38.2	58.1	42.3	38.3	44.2
BlendMask-baseline	Swin-base	4.7	44.4	65.2	48.2	42.1	50.9
TPR-BlendMask	Swin-tiny	10.2	40.0	62.0	43.4	38.9	46.7
TPR-BlendMask	Swin-base	3.5	45.9	67.3	51.2	43.6	52.3

TPR improves the results with considerable margins for different approaches. Note that all the methods are implemented in the same framework under the same setting for fair comparison.

Computation Analysis Since our method is dynamic and it has different GFlops for different inputs. We report several different settings in Tab. 1(f) including the maximum GFlops, average GFlops and minimum GFlops. The results prove that our TPR can not only have fewer computational overhead but also improve performance by a large margin. For instance, our method obtains about 2.1%-2.5% relative mAP gains over the static baseline with a lower average computational complexity (relatively only about 6.4% - 7.7%). That indicates the potential of our approach for the application purpose.

Visualization of Gate Maps In Fig. 6, we visualize several examples of our double gates. We observe that the inner gates mainly provide detailed and fine-grained instance details from the reference frame (such as heads or foot on beers) while the outer-gates focus on the current foreground objects (roughly location of instances in current frame). Both gates prohibit background noises and make the inference more efficient, which is consistent with our motivation in Sec. 1.

Qualitative Results In Fig. 7, we give two visual examples of the baseline method and our TPR with each group containing images sampled from the same video. The same color represents the same instance identity. From these results, we observe that our TPR can segment instances well in two challenging situations: (1) instance overlapping (The first group , two deers are very close to each other), (2) instance deformations and large variations (The second groups, the appearance of the cow has changed greatly over time, thus the cow is classified into the dog wrongly.). Based on these results, we observe that our TPR reliably propagates multi-scale features that are specific to each instance with object deformations and large variations (The second example, the cow is classified into dog) and overlapping (The first example, two deers are very close with each other), in object appearance.

Comparison with the Previous Works on YouTube-VIS-2019

Finally, we give a detailed comparison on our methods with previous work in Tab. 2. The comparison contains several aspects,

TABLE 3: **Video instance segmentation AP (%)** on the new YouTube-VIS-2021 [1] validation dataset.

Method	backbone	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
MaskTrack R-CNN [1]	ResNet50	30.3	51.1	32.6	31.0	35.5
SipMask-VIS [24]	ResNet50	31.7	52.5	34.0	30.8	37.8
BlendMask-VIS [23]	ResNet50	32.5	53.3	35.0	31.8	38.8
Cross-VIS [57]	ResNet50	34.2	54.4	37.9	30.4	38.2
TPR-BlendMask	ResNet50	35.3	54.8	37.9	34.2	43.6
TPR-BlendMask	ResNet101	38.1	59.1	43.4	38.8	45.6

TABLE 4: Comparisons on Cityscapes-VPS dataset. We report our reproduced baseline in [2].

Method	VPQ	<i>V PQ_{thing}</i>	<i>V PQ_{stuff}</i>
UPSNet+Track [1] +TPR	55.8 56.7	43.9 44.5	65.3 66.1
UPSNet+Track+Fuse [2] +TPR	56.8 57.5	44.2 44.7	66.4 66.8

including accuracy, speed, inference type and backbone network. Though MaskProp [3] obtains higher mAP, that method is much slower. Moreover, MaskProp combines multiple networks such as video object detection network [52], Hybrid Task Cascade Network [5] for cascade feature learning and the complex High-Resolution Mask Refinement post-processing. Our method shares a much simpler pipeline and performs the online inference. As shown in Tab. 2, our TPR improves the various models with little FPS drop. In particular, with ResNet101 and DCN [49] as backbone, our method can achieve **42.2% mAP** while running at 8.2 FPS. We also verify the generality on larger Swin Transformer [26] as backbone in the bottom of Tab. 2. After using the Swin-base backbone, our method can achieve 45.9 AP while running at 4.8 FPS which is better and faster than the MaskProp.

Comparison with the Previous Works on YouTube-VIS-2021

We report the recently proposed MaskTrack R-CNN and SipMask-VIS on this dataset using official implementation for comparison. We also compare the recently proposed CrossVIS [57] for fair comparison. From the Tab. 3, our method achieves better results than previous works.

4.3 Experiments on Cityscapes-VPS for VPS

In Tab. 4, we compare our methods with strong UPSNet with track head as baseline on Cityscapes-VPS dataset. Our method can achieve significant gain on both thing class and stuff class. This proves the potential of TPR for other video scene understanding tasks.

5 CONCLUSION

In this paper, we propose a conceptually new framework named Temporal Pyramid Routing (TPR) for dynamic instance learning in video. The pyramid features are well aligned and routed along the temporal dimension via TPR. In particular, we design specific Dynamic Aligned Cell Routing (DPCR) with a double gate design to avoid background noise while capturing useful information during the propagation. Moreover, the Cross Pyramid Routing (CPR) is proposed to propagate aligned features into all scales. Extensive experiments demonstrate the effectiveness and efficiency of our TPR on VIS and VPS task. We believe this work can provide several insights into future works for modeling dynamics in video in the aspect of video dynamic network design.

REFERENCES

- [1] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *ICCV*, 2019.
- [2] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *CVPR*, 2020.
- [3] G. Bertasius and L. Torresani, "Classifying, segmenting, and tracking object instances in video with mask propagation," in *CVPR*, 2020.
- [4] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *CVPR*, 2021.
- [5] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *CVPR*, 2019.
- [6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.
- [7] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, and B. Leibe, "Stem-seg: Spatio-temporal embeddings for instance segmentation in videos," in *ECCV*, 2020.
- [8] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool, "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth," in *CVPR*, 2019.
- [9] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
- [10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.
- [11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.
- [12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CVPR*, Jun 2018.
- [13] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *CVPR*, 2019.
- [14] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *CVPR*, 2020.
- [15] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, 2017.
- [16] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *ICCV*, 2017.
- [17] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *CVPR*, 2020.
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015.
- [19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.
- [20] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *CVPR*, 2018.
- [21] L. Song, Y. Li, Z. Jiang, Z. Li, H. Sun, J. Sun, and N. Zheng, "Fine-grained dynamic head for object detection," in *NIPS*, 2020.
- [22] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," in *CVPR*, 2020.
- [23] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-down meets bottom-up for instance segmentation," in *CVPR*, 2020.
- [24] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," *ECCV*, 2020.
- [25] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *ICCV*, 2019.
- [26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [27] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *ECCV*. Springer, 2016.
- [28] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*. Springer, 2014.
- [29] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *CVPR*, 2019.
- [30] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*, 2020.
- [31] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," *arXiv preprint arXiv:1708.02551*, 2017.
- [32] C.-C. Lin, Y. Hung, R. Feris, and L. He, "Video instance segmentation tracking with a modified vae architecture," in *CVPR*, 2020.
- [33] Y. Fu, L. Yang, D. Liu, T. S. Huang, and H. Shi, "Comfeat: Comprehensive feature aggregation for video instance segmentation," *AAAI*, 2021.
- [34] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *CVPR*, 2019.
- [35] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, "Video k-net: A simple, strong, and unified baseline for video segmentation," in *CVPR*, 2022.
- [36] H. Yuan, X. Li, Y. Yang, G. Cheng, J. Zhang, Y. Tong, L. Zhang, and D. Tao, "Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation," 2021.
- [37] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "Transvod: End-to-end video object detection with spatial-temporal transformers," 2022.
- [38] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *CVPR*, 2020, pp. 591–600.
- [39] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *NIPS*, 2019.
- [40] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *CVPR*, 2020.
- [41] A. Almahairi, N. Ballas, T. Coopmans, Y. Zheng, H. Larochelle, and A. Courville, "Dynamic capacity networks," in *ICML*, 2016.
- [42] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *ICLR*, 2018.
- [43] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," *ICLR*, 2019.
- [44] H. Cai, L. Zhu, and S. Han, "Proxylessnas: Direct neural architecture search on target task and hardware," *arXiv preprint arXiv:1812.00332*, 2018.
- [45] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *ICML*, 2018.
- [46] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.
- [47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.
- [48] Z. Tian, C. Shen, H. Chen, and T. He, "Fcose: A simple and strong anchor-free object detector," *TPAMI*, 2021.
- [49] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017.
- [50] B. Graham, "Spatially-sparse convolutional neural networks," *Computer Science*, vol. 34, no. 6, 2014.
- [51] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," *ECCV*, 2016.
- [52] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *ECCV*, 2018.
- [53] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "Upsnet: A unified panoptic segmentation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8818–8826.
- [54] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *CVPR*, 2019.
- [55] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *CVPR*, 2015.
- [56] H. Liu, R. A. R. Soto, F. Xiao, and Y. J. Lee, "Yolactedge: Real-time instance segmentation on the edge (jetson agx xavier: 30 fps, rtx 2080 ti: 170 fps)," *arXiv preprint arXiv:2012.12259*, 2020.
- [57] S. Yang, Y. Fang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Crossover learning for fast online video instance segmentation," *arXiv preprint arXiv:2104.05970*, 2021.
- [58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.
- [60] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," <https://github.com/facebookresearch/detectron2>, 2019.
- [61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [62] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*, 2017.
- [63] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, "Feelvos: Fast end-to-end embedding learning for video object segmentation," in *CVPR*, 2019.

- [64] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, “Efficient video object segmentation via network modulation,” in *CVPR*, 2018.