# MED-VT: Multiscale Encoder-Decoder Video Transformer with Application to Object Segmentation

Rezaul Karim He Zhao Richard P. Wildes Mennatullah Siam
York University
{karimr31, zhufl, msiam}@eecs.yorku.ca, wildes@cse.yorku.ca

## Abstract

*Multiscale video transformers have been explored in a wide variety of vision tasks. To date, however, the multiscale processing has been confined to the encoder or decoder alone. We present a unified multiscale encoder-decoder transformer that is focused on dense prediction tasks in videos. Multiscale representation at both encoder and decoder yields key benefits of implicit extraction of spatiotemporal features (*i.e. *without reliance on input optical flow) as well as temporal consistency at encoding and coarse-to-fine detection for high-level (*e.g. *object) semantics to guide precise localization at decoding. Moreover, we propose a transductive learning scheme through many-to-many label propagation to provide temporally consistent predictions. We showcase our Multiscale Encoder-Decoder Video Transformer (MED-VT) on Automatic Video Object Segmentation (AVOS) and actor/action segmentation, where we outperform state-of-the-art approaches on multiple benchmarks using only raw images, without using optical flow.*

## 1. Introduction

Transformers have been applied to a wide range of image and video understanding tasks as well as other areas [13]. The ability of such architectures to establish data relationships across space and time without the local biases inherent in convolutional and other similarly constrained approaches arguably is key to the success. Multiscale processing has potential to enhance further the learning abilities of transformers through cross-scale learning [4, 6, 8, 19, 43]. A gap exists, however, as no approach has emerged that makes full use of multiscale processing during both encoding and decoding in video transformers. Recent work has focused on multiscale transformer encoding [8, 19], yet does not incorporate multiscale processing in the transformer decoder. Other work has proposed multiscale transformer decoding [6], yet was designed mainly for single images and did not consider the structured prediction nature inherent to video tasks, *i.e.* the importance of temporal consistency.

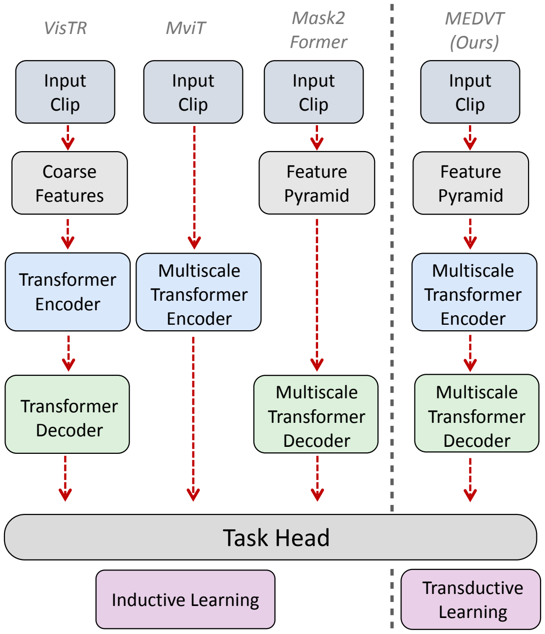

Figure 1. Comparison of state-of-the-art multiscale video transformers and our approach. Video transformers take an input clip and feed features to a transformer encoder-decoder, with alternative approaches using multiscale processing only in the encoder or decoder. We present a unified multiscale encoder-decoder video transformer (MED-VT), while predicting temporally consistent segmentations through transductive learning. We showcase MED-VT on two tasks, video object and actor/action segmentation.

In response, we present the first Multiscale Encoder Decoder Video Transformer (MED-VT). At encoding, its within and between-scale attention mechanisms allow it to capture both spatial, temporal and integrated spatiotemporal information. At decoding, it introduces learnable coarse-to-fine queries that allow for precise target delineation, while enforcing temporal consistency among the predicted masks through transductive learning [38, 52].

We primarily illustrate the utility of MED-VT on the task of Automatic Video Object Segmentation (AVOS). AVOS separates primary foreground object(s) from background in a video without any supervision, *i.e.* without information on the objects of interest [53]. This task is important and challenging, as it is a key enabler of many subsequent visually-guided operations, *e.g*., autonomous driving and augmented reality. AVOS shares challenges common to any VOS task (*e.g*., object deformation and clutter). Notably, however, the requirement of complete automaticity imposes extra challenges to AVOS, as it does not benefit from any per video initialization. Lacking prior information, solutions must exploit appearance (*e.g*., colour and shape) as well as motion to garner as much information as possible.

MED-VT responds to these challenges. Its within and between scale attention mechanisms capture both appearance and motion information as well as yield temporal consistency. Its learnable coarse-to-fine queries allow semantically laden information at deeper layers to guide finer scale features for precise object delineation. Its transductive learning through many-to-many label propagation ensures temporally consistent predictions. To showcase our model beyond AVOS, we also apply it to actor/action segmentation. Figure 1 overviews MED-VT compared to others.

## 2. Related work

**Video dense prediction tasks.** We focus on two important aspects of video dense prediction: the multiscale nature of objects and temporally consistent spatial localization for per pixel classification, as well as operation without the expense of optical flow. For tasks, we consider two dense prediction tasks, Automatic Video Object Segmentation (AVOS) [53] and actor/action segmentation, while leaving extensions to instance-aware AVOS [24, 40] and tracking [28, 45] for future work. Dominant approaches to AVOS rely on both colour images and optical flow as input [10, 32, 34, 54]. Other approaches consider attention [23, 41] to capture recurring objects in a video via simple mechanisms, *e.g*. co-attention. Similarly, dominant approaches to actor/action segmentation depend on optical flow [7, 12].

**Multiscale processing.** Multiscale processing is an established technique across computer vision. Some representative examples in the era of convolutional networks include edge detection [46], image segmentation [33], object detection [20] as well as AVOS [10, 32, 54]. Recently, multiscale processing has been applied with transformers to assist the understanding of single images (*e.g*. classification [43], detection [4, 55] and panoptic segmentation [6]) and videos (*e.g*. action recognition [8, 19]). However, such transformers are limited by lack of unified multiscale processing (*i.e*. restricted to the encoding phase) or not readily applicable to video understanding (*i.e*. primarily used for static images [43]), in general, and dense video predictions, in particular. In contrast, while our work exploits multiscale information, it makes fuller use in its multiscale encoder-decoder via *Within* and *Between* scale attention.

**Temporal consistency.** AVOS models typically benefit from leveraging the principle of global consistency across multiple frames. Early efforts sought such consistency on the feature level by fusing the appearance (*e.g*. RGB images) and motion information (*e.g*. optical flow) of given videos [10, 32, 54]. However, these required additional effort on high-quality flow estimation. Other work focused on enforcing consistency between features computed across time using co-attention [23]. A limitation of this approach is its excessive computational overhead, because of its need for multiple inference iterations to yield good results. Recent advances in semi-automatic VOS have devised a lightweight yet efficient counterpart: A prediction-level label propagator that explicitly exploits frame-wise semantic consistency, which was proposed in a transductive inference setting [25, 50]. Nonetheless, their propagator was confined to a single frame at a time. We present a label propagator that extends the existing approach by considering many-to-many propagation, to effectively capture temporal dependency within an entire input clip for AVOS.

**Contributions.** In the light of previous work, our contributions are threefold. (1) We present the first end-to-end multiscale transformer for video understanding dense prediction relying solely on raw images without optical flow input. (2) Our model is the first to intergrate multiscale transformer encoder and decoder in any video understanding task. The encoder enables our model to capture spatiotemporal information across scales, while the multiscale decoder provides precise localization. (3) We present the first many-to-many label propagation scheme within a transductive learning paradigm to ensure temporally consistent predictions across an entire input clip. Our approach outperforms the state of the art on multiple AVOS and actor/action segmentation datasets. Our code is available at: rkyuca.github.io/medvt.

## 3. MED-Video Transformer (MED-VT)

### 3.1. Overview

MED-VT is an end-to-end video transformer that inputs a clip and provides dense segmentation predictions without the need of explicit optical flow. Processing unfolds in five main stages; see Fig. 2. (i) A feature pyramid is extracted using a backbone network. (ii) The extracted stacks of feature pyramids are processed by a transformer encoder and (iii) decoder. (iv) A task specific head produces initial predictions. (v) A many-to-many temporal label propagator refines the initial predictions by enforcing temporal consistency. Our architecture is unique in its unified approach to encoding and decoding at multiple scales, as well as its use of many-to-many label propagation.

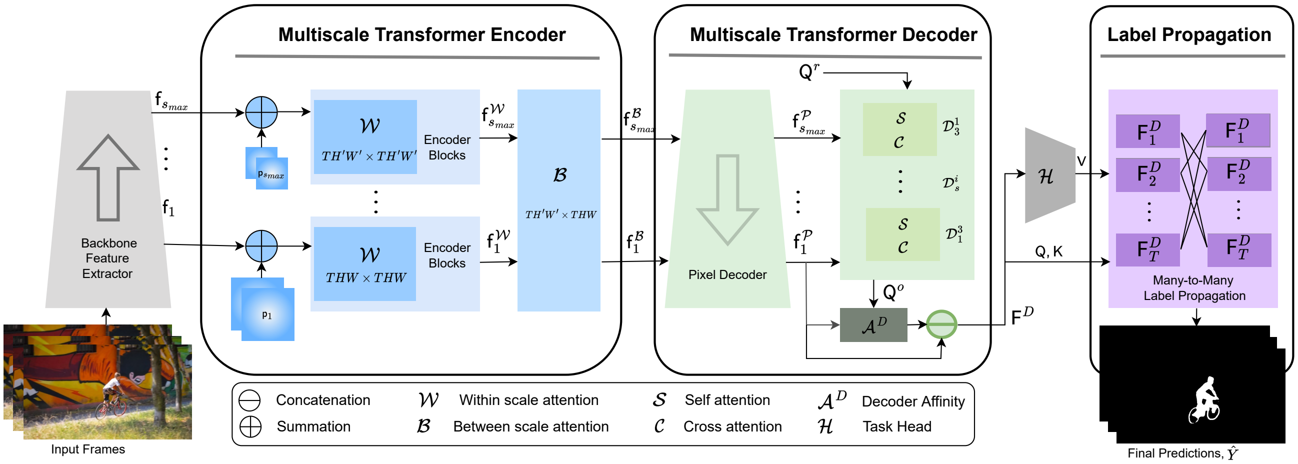

Figure 2. Detailed (MED-VT) architecture with unified multiscale encoder-decoder transformer, illustrated with application to Automatic Video Object Segmentation (AVOS). The model has five functionally distinct components. (i) Backbone feature extractor to extract per frame features, $f_s$, at multiple scales, $s \in \{1, \cdots, s_{max}\}$. (ii) Multiscale transformer encoder consisting of spatiotemporal within and between scale attention with resulting features, $f^{\mathcal{W}}_s$ and $f^{\mathcal{B}}_s$, resp; the multihead attention transformation, (1b), is used for both. (iii) Multiscale transformer decoder consisting of pixel decoding, which produces decoded features, $f^{\mathcal{P}}_s$, and a series of mulitscale query learning decoder blocks, $\mathcal{D}^i_s$, for the corresponding $i^{th}$ iteration and scale $s$, each of which entail self and cross attention, again using the multihead attention transformation, (1b). The input to the blocks are the decoded features $f^{\mathcal{P}}_s$ and the query resulting from the previous block, with a randomized query, $Q^r$, initialization; the output is a final object query, $Q^o$. The decoder applies an affinity, (6), between $Q^o$ and the finest scale decoded features, $f^{\mathcal{P}}_1$, to yield an object attention map, which is concatenated with the finest scale decoded features for final decoder output, $F^D$. (iv) A task specific head, $\mathcal{H}$, that inputs $F^D$ to produce initial predictions. (v) Many-to-many label propagation that inputs the initial predictions as values, V, as well as $F^D$ as queries, Q, and keys, K, to yield temporally consistent segmentation final masks, $\hat{Y}$. Our key innovations, outlined in bold boxes, lie in the unified multiscale encoder-decoder and label propagator.

Feature extraction is standard. Given a video clip, we first extract a set of multiscale features, $F = \{f_s : s \in S\}$, where $f_s \in \mathbb{R}^{T \times H_s \times W_s \times C_s}$ represent features extracted at scale $s$, $S = \{1, .., s_{max}\}$ indexes the scale stages from fine to coarse and $\{T, H_s, W_s, C_s\}$ are the clip length, height, width and channel dimension at scale $s$, resp. Prior to subsequent processing, backbone features are down projected to $d$ dimensions, $\bar{f}_s = \phi(f_s) \in \mathbb{R}^{TH_sW_s \times d}$, where $\phi$ is a simple $1 \times 1$ convolutional layer, followed by flattening. Our model is not specific to a particular backbone feature extractor; indeed, we illustrate with multiple in Sec. 5. Moreover, our model is not specific to a particular task head and we also illustrate with multiple in Sec. 5. The rest of this section, details our novel encoder-decoder and label propagator.

### 3.2. Multiscale transformer encoder

Transformer encoders built on spatiotemporal self-attention mechanisms can capture long range object representation relationships across both space and time for video recognition tasks [2, 3, 8]. They thereby naturally support learning of both spatial and temporal features as well as integrated spatiotemporal features. Notably, however, standard encoders that operate over only coarse scale feature maps limit the ability to capture fine grained pattern structure as well as fail to support precise localization [44]. To overcome these limitations, our encoder encompasses two main operations of within and between-scale spatiotemporal attention on multiple feature abstraction levels with different resolutions that are extracted from a backbone convolutional network, as discussed in Sec. 3.1.

We formulate the operations of within and between scale attention via standard multihead attention, $\mathcal{M}$, defined as

$$\mathcal{A}_h(Q, K, V) = \text{Softmax}\left(\frac{1}{\sqrt{d}} QW^q_h (KW^k_h)^\top\right) VW^v_h, \quad (1a)$$

$$\mathcal{M}(Q, K, V) = \text{Concat}^{N_h}_{h=1}(\mathcal{A}_h(Q, K, V))W^o, \quad (1b)$$

where Q, K and V are query, key and value, resp., while $W^q_h$, $W^k_h$ and $W^v_h$ are their corresponding learned weight matrices for head $h$, $d$ is feature dimension and $W^o$ is the weight matrix for the final multiheaded output.

**Within scale attention.** We formulate multihead $\mathcal{W}$ithin scale attention by instantiating multihead attention, (1b), as

$$\mathcal{W}(\bar{f}_s, p_s) = \mathcal{M}(\bar{f}_s + p_s, \bar{f}_s + p_s, \bar{f}_s), \quad (2)$$

with $p_s \in \mathbb{R}^{TH_sW_s \times d}$ per scale positional encodings to preserve location information, *cf*. [44]. $\mathcal{W}$ is applied successively across multiple encoder layers. We compute the final encoded feature maps, $F^{\mathcal{W}} = \{\bar{f}^{\mathcal{W}}_s : s \in S\}$, for each corresponding scale, to capture globally consistent representation of objects of interest. Successive application of spatiotemporal within scale attention yields globally coherent representation, otherwise limited by local convolutions.

**Between scale attention.** For $\mathcal{B}$etween-scale attention, $\mathcal{B}$, we apply attention on the encoded features, $F^{\mathcal{W}}$. Coarse scale feature maps capture rich semantics by virtue of having gone through multiple abstraction layers. Correspondingly, the feature map from a coarser scale, $s$, *i.e.* $\bar{\mathsf{f}}^{\mathcal{W}}_s$, is used to affect the immediately finer scale feature map, $\bar{\mathsf{f}}^{\mathcal{W}}_{s-1}$, based on between-scale attention affinity. To achieve this goal, we again use multihead attention, (1b), now as

$$\mathcal{B}(\bar{\mathsf{f}}^{\mathcal{W}}_{s-1}, \mathsf{p}_{s-1}, \bar{\mathsf{f}}^{\mathcal{W}}_s, \mathsf{p}_s) = \mathcal{M}(\bar{\mathsf{f}}^{\mathcal{W}}_{s-1} + \mathsf{p}_{s-1}, \bar{\mathsf{f}}^{\mathcal{W}}_s + \mathsf{p}_s, \bar{\mathsf{f}}^{\mathcal{W}}_s) \tag{3}$$

where $\mathsf{p}_s$ and $\mathsf{p}_{s-1}$ are positional embeddings. This operation enhances between-scale communication to promote globally consistent, semantically rich feature maps and is conducted between each pair of adjacent scales. We denote the output features from between-scale attention, $\mathcal{B}$, as $F^{\mathcal{B}} = \{\bar{\mathsf{f}}^{\mathcal{B}}_s : s \in S\}$.

### 3.3. Multiscale transformer decoder

Our multiscale encoder's between-scale attention promotes spatiotemporal consistency across scales, while the decoder promotes multiscale query learning to localize object-level properties. Our decoder works in two steps: (i) pixel decoding, which propagates coarse scale semantics to fine scale localization and (ii) transformer decoding, which generates adaptive queries.

**Pixel decoding.** In pixel decoding we seek to propagate semantics of coarse scale features to finer scales. For this purpose, we use a Feature Pyramid Network (FPN) [20]. The FPN works top down from coarse features with highest abstraction to fine features by injecting coarser scale information into each finer scale. It thereby allows for better communication from high level to low level semantics with finer details preserved before queries are generated in the actual transformer decoder. The FPN inputs the between-scale attention features, $F^{\mathcal{B}}$, and outputs a feature pyramid $F^{\mathcal{P}} = \{\bar{\mathsf{f}}^{\mathcal{P}}_s : s \in S\}$. See supplement for details.

**Decoding and adaptive queries.** Improved object queries, $\mathsf{Q}^o$, are learned via multiscale coarse-to-fine processing, where the queries adapt to the input features for object localization. This multiscale processing enriches the learnable queries so they work better on the finest resolution. These adaptive queries, $\mathsf{Q}^o \in \mathbb{R}^{N_q \times d}$, with $N_q$ the number of queries, are learned jointly from the multiscale feature maps, $F^{\mathcal{P}}$, using a series of transformer decoder blocks, $\mathcal{D}^i_s$ operating coarse-to-fine across scales, $s$, and with multiple iterations, $i$; see Fig. 3. Each transformer decoder block, $\mathcal{D}^i_s$, inputs pixel decoded features, $\bar{\mathsf{f}}^{\mathcal{P}}_s$, at a particular scale, $s$. At each iteration, $i$, the blocks operate, coarse-to-fine, with queries output from the previous serving as input (along with decoded features from $F^{\mathcal{P}}$) to the next. The process iterates $N_d$ times, as notated by superscripts $i$ on the blocks, $\mathcal{D}^i_s$, *i.e.* $i \in \{1...N_d\}$. The entire process starts with a randomly initialized query, $\mathsf{Q}^r$ and culminates in the final adaptive object queries, $\mathsf{Q}^o$. The adaptive queries serve to compactly represent the foreground object with its different appearance changes and deformation within the input clip.

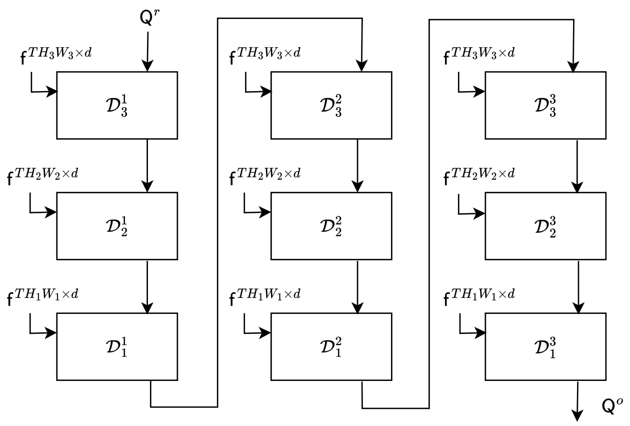
$\mathsf{Q}^r$
$\mathsf{f}^{TH_3W_3\times d}$
$\mathcal{D}^1_3$
$\mathsf{f}^{TH_2W_2\times d}$
$\mathcal{D}^1_2$
$\mathsf{f}^{TH_1W_1\times d}$
$\mathcal{D}^1_1$
$\mathsf{f}^{TH_3W_3\times d}$
$\mathcal{D}^2_3$
$\mathsf{f}^{TH_2W_2\times d}$
$\mathcal{D}^2_2$
$\mathsf{f}^{TH_1W_1\times d}$
$\mathcal{D}^2_1$
$\mathsf{f}^{TH_3W_3\times d}$
$\mathcal{D}^3_3$
$\mathsf{f}^{TH_2W_2\times d}$
$\mathcal{D}^3_2$
$\mathsf{f}^{TH_1W_1\times d}$
$\mathcal{D}^3_1$
$\mathsf{Q}^o$

Figure 3. The decoder stacked coarse-to-fine processing. Our multiscale decoder inputs a multiscale feature pyramid, $F^{\mathcal{P}}$, and randomly initialized queries, $\mathsf{Q}^r$, and outputs final object queries, $\mathsf{Q}^o$. The input is processed coarse-to-fine and iteratively through multiple decoder blocks, $\mathcal{D}^i_s$, with $s$ indicating input feature scale and $i$ indicating iteration. For simplicity, we show $s = 3$ scales and $i = 3$ iterations, with $\mathsf{f}^{-}$ denoting features from each level of the pyramid, $F^{\mathcal{P}}$, where corresponding dimensions of the three levels are, $TH_3W_3 \times d, TH_2W_2 \times d, TH_1W_1 \times d$, resp.

In each decoder block both self and cross attention operates, as in encoding attention, Sec. 3.2. To define $\mathcal{S}$elf attention, we instantiate multihead attention, (1b), as

$$\mathcal{S}(\mathsf{Q}^i_s, \mathsf{p}^{\mathsf{Q}}_s) = \mathcal{M}(\mathsf{Q}^i_s + \mathsf{p}^{\mathsf{Q}}_s, \mathsf{Q}^i_s + \mathsf{p}^{\mathsf{Q}}_s, \mathsf{Q}^i_s), \tag{4}$$

with $\mathsf{Q}^i_s$ the input query to block $\mathcal{D}^i_s$ and $p^{\mathsf{Q}}_s \in \mathbb{R}^{N_q \times d}$ learnable *query positional embeddings*. To define $\mathcal{C}$ross attention, we instead instantiate (1b) as

$$\mathcal{C}(\mathsf{Q}^i_s, \mathsf{p}^{\mathsf{Q}}_s, \bar{\mathsf{f}}^{\mathcal{P}}_s, \mathsf{p}_s, \hat{\mathsf{p}}^{\sigma}_s) = \mathcal{M}(\mathsf{Q}^i_s + \mathsf{p}^{\mathsf{Q}}_s, \bar{\mathsf{f}}^{\mathcal{P}}_s + \mathsf{p}_s + \hat{\mathsf{p}}^{\sigma}_s, \bar{\mathsf{f}}^{\mathcal{P}}_s), \tag{5}$$

with $\mathsf{p}_s \in \mathbb{R}^{TH_sW_s \times d}$ *feature positional embeddings* and $\hat{\mathsf{p}}^{\sigma}_s \in \mathbb{R}^{TH_sW_s \times d}$ derived from learnable scale embeddings, $\mathsf{p}^{\sigma}_s \in \mathbb{R}^{1 \times d}$, after being repeated across $T, H_s, W_s$; this operation allows cross attention to be scale sensitive, *cf*. [6].

After all decoder blocks have produced the query, $\mathsf{Q}^o$, a final cross attention block is used to establish affinity between the query and finest scale features, $\bar{\mathsf{f}}^{\mathcal{P}}_1$, and thereby generate an object attention map, $\mathsf{F}^{\mathcal{A}}$. Since only the query and features are considered, a two argument affinity is used,

$$
\begin{aligned}
\mathsf{F}^{\mathcal{A}} &= \mathcal{A}^{D}(\mathsf{Q},\mathsf{K}) \\
&= \mathrm{Concat}_{h=1}^{N_h}\left[\mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\mathsf{Q}\mathsf{W}_h^{q}(\mathsf{K}\mathsf{W}_h^{k})^{\top}\right)\right]
\end{aligned}
\tag{6}
$$

with $\mathsf{Q} = \mathsf{Q}^{o}, \mathsf{K} = \bar{\mathsf{f}}_1^{\mathcal{P}}$, $N_h$ the number of heads and $\mathsf{W}_h^{q}$, $\mathsf{W}_h^{k}$ being learnable parameters for head $h$. The final decoder output, $\mathsf{F}^{D}$, is formed as the concatenation of the finest scale features, $\mathsf{f}_1^{\mathcal{P}}$, with the attention maps, $\mathsf{F}^{\mathcal{A}}$, *i.e.* $\mathsf{F}^{D} = \mathsf{F}^{\mathcal{A}} \ominus \mathsf{f}_1^{\mathcal{P}}$, with $\ominus$ channel-wise concatenation. This augmentation further enhances localization precision.

### 3.4. Many-to-many temporal label propagation

Label propagation is a standard technique that can be used in transductive reasoning [52]. In semi-supervised VOS, it was proposed to train an end-to-end model to propagate the labels from many/all previous frames, $\{\cdots, t-1\}$, to a single current frame, $t$, hence causal many-to-one propagation within a transductive setting [25]. This operation provides structured prediction across frames instead of independent predictions per frame. We extend this idea by allowing label propagation from all other frames, $\{\cdots, t-1, t+1, \cdots\}$, in a clip to each frame, $t$, hence many-to-many label propagation. This extended operation enforces structured prediction across all frames in a clip.

Our many-to-many label propagator three operators, sequentially applied: (i) a label encoder, $\mathcal{E}_L$, (ii) a spatiotemporal affinity based label propagator using masked attention, $\mathcal{M}^{m}$, and (iii) a label decoder, $\mathcal{D}_L$. The input to the encoder is an initial prediction, $\mathsf{Y}' = \mathcal{H}(\mathsf{F}^{D})$, generated by a task head, $\mathcal{H}$, from the output of the previous decoding, $\mathsf{F}^{D}$. The label encoder then generates an encoding of dimension $D$ from the initial predictions, $\mathsf{Y}'$, and flattens it, $\bar{\mathsf{Y}} = \mathcal{E}_L(\mathsf{Y}') \in \mathbb{R}^{TH_1W_1 \times D}$, *cf*. [25]. The label propagator extends the encoded labels temporally in a many-to-many fashion. The label decoder, $\mathcal{D}_L$, takes these propagated encoded labels and generates the final class-wise predictions. The label encoder, $\mathcal{E}_L$, is a similar CNN to that used elsewhere [25]. The label decoder, $\mathcal{D}_L$, is a three-layer CNN.

We devise the label propagator as a masked attention module [36, 39] to capture the long-distance dependencies between labels while preserving efficiency. The mask, $\mathsf{M} \in \mathbb{R}^{TH_1W_1 \times TH_1W_1}$, restricts attention to regions centered around the predicted data point, akin to the notion of *clique* in conditional random fields [16] or, more generally, graph theory. The mask can be defined to promote communication between data points in a wide variety of fashions (*e.g.* within frame, between frames, many-to-one, many-to-many) [16]. We use this mechanism for temporal many-to-many propagation to encourage information sharing among different frames. The mask of two arbitrary positions is set to, $\mathsf{M}_{ij} = -\infty$, if they are in the same frame, otherwise as zero. Formally, the mask is defined as

$$
\mathsf{M}_{ij} = \begin{cases} -\infty, & \text{if } \tau(i) = \tau(j) \\ 0, & \text{otherwise.} \end{cases}
\tag{7}
$$

where $\tau(\cdot)$ returns the frame index of given data points. Masked attention is then defined by augmenting the standard multihead attention, (1b), to become

$$
\begin{aligned}
&\mathcal{A}_h^{m}(\mathsf{Q},\mathsf{K},\mathsf{V},\mathsf{M}) = \\
&\quad \mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\mathsf{Q}\mathsf{W}_h^{q}(\mathsf{K}\mathsf{W}_h^{k})^{\top} + \mathsf{M}\right)\mathsf{V}\mathsf{W}_h^{v},
\end{aligned}
\tag{8a}
$$

$$
\mathcal{M}^{m}(\mathsf{Q},\mathsf{K},\mathsf{V},\mathsf{M}) = \mathrm{Concat}_{h=1}^{N_h}(\mathcal{A}_h^{m}(\mathsf{Q},\mathsf{K},\mathsf{V},\mathsf{M}))\mathsf{W}^{o},
\tag{8b}
$$

where $\mathsf{Q},\mathsf{K},\mathsf{V},\mathsf{M}$ are the queries, keys, values and mask resp., while $\mathsf{W}_h^{q}$, $\mathsf{W}_h^{k}$, $\mathsf{W}_h^{v}$ are learnable parameters for head $h$ and $W^{o}$ is the final weighting for combining heads. Our label propagator instantiates masked attention, (8b), as

$$
\tilde{\mathsf{Y}} = \mathcal{M}^{m}(\bar{\mathsf{F}}^{D}, \bar{\mathsf{F}}^{D}, \bar{\mathsf{Y}}, \mathsf{M}),
\tag{9}
$$

where $\bar{\mathsf{F}}^{D} \in \mathbb{R}^{TH_1W_1 \times (d+N_h)}$ is the flattened decoded features. As defined, this operation propagates labels across all data points in the entire clip, both spatial position and frames, unlike previous efforts that were limited to propagating to the current frame only [25].

Finally, the overall class-wise predictions, $\hat{\mathsf{Y}}$, are produced by combining the propagated label decodings, $\mathcal{D}_L(\tilde{\mathsf{Y}})$, and the initial predictions from the segmentation head, $\mathsf{Y}'$ according to

$$
\hat{\mathsf{Y}} = \frac{1}{2}(\mathcal{D}_L(\tilde{\mathsf{Y}}) + \mathsf{Y}').
\tag{10}
$$

We combine the initial predictions per frame and the propagated predictions from all other frames because while label propagation enforces temporal consistency, it also can sacrifice boundary precision due to the smoothing it incurs. The final combination, (10), provides both temporal consistency and precise boundaries. The supplement explores theoretical connections between our label propagation approach and spectral clustering.

## 4. Learning scheme

For both the AVOS and actor/action segmentation cases, we train the model using a combination of distribution- and region-based losses. In particular, we combine the distribution-based focal [21] and the region-based Dice [26] losses. Notably, this combination ameliorates the challenge of class imbalance [21,26], as in both our tasks, background pixels are more frequent than other classes. To support many-to-many label propagation in AVOS we use the entire clip groundtruth to compute the loss. However, for the actor/action segmentation dataset, we only have a groundtruth label for the centre frame; so, it is not directly applicable to perform many-to-many label propagation. Thus, in actor/action segmentation we initially train our model without label propagation and compute pseudo-labels for unlabelled frames. Then the groundtruth of the centered frame and pseudolabels for the rest of the frames within the clip are used to train our model with label propagation to enforce the many-to-many label consistency across the clip. Further details on the learning scheme are in the supplement.

# 5. Empirical evaluation

## 5.1. Experiment design

**Implementation details.** We present results from two backbones, ResNet-101 [9] and Video-Swin [22], to extract multiscale feature pyramid, $F$, from a clip of $T = 6$ frames by taking features, $\mathsf{f}_s$, from successive stages. For the multiscale transformer encoder, we use it on the two coarsest scale features for memory efficiency reasons; we use six and one encoder blocks for the two smallest scale features. In the transformer decoder, we use three feature scales and three iterations, $N_d = 3$, resulting in nine decoding layers. All multihead attention operations and the attention block, $\mathcal{A}^D$, have, $N_h = 8$, attention heads and use a channel of dimension $d = 384$. The final segmentation head, $\mathcal{H}$, is a three-layer convolutional module. The final decoding layer of $\mathcal{H}$ is defined according to the segmentation task to match the number of categories, i.e., foreground vs. background for AVOS and 43 to encompass the classes in the actor/action dataset. Additional architecture and training details are provided in the supplement.

**Inference, datasets and evaluation protocols.** For inference we use the same clip length as in training, $T = 6$, in a sliding window with the predicted logits upsampled to the original image size. Each temporal window of frames serves as input to our model to predict the segmentation of its centre frame. For the DAVIS'16 dataset only, we use multiscale inference postprocessing wherein inference is conducted at multiple scales and subsequently averaged [5], as it is standard with that dataset to present results with and without postprocessing; although, the postprocessing methods vary. The supplement details our multiscale inference postprocessing.

For AVOS, we test on three standard datasets: DAVIS'16 [29], YouTube-Objects [30] and MoCA (Moving Camouflaged Animals) [17]. DAVIS'16 is a widely adopted AVOS benchmark, while YouTube-Objects is another large-scale VOS dataset. MoCA is the most challenging motion segmentation dataset available, as in the absence of motion the camouflaged animals are almost indistinguishable from the background by appearance alone (*i.e.* colour and texture). For actor/action segmentation, we use the A2D dataset [47]. For all datasets we use its standard evaluation protocol. Further dataset details are provided in the supplement.

## 5.2. Comparison to the state-of-the-art

**MoCA.** Table 1 shows MoCA results, with comparison to the previous state of the art. Since the dataset provides only bounding box annotations, following standard protocol, we compare maximum bounding box of our segmentation mask to compute region similarity. It is evident that MED-VT outperforms all others by a notable margin when using the same backbone (*i.e.* ResNet-101) as the previous state of the art [32, 54]. Moreover, MED-VT performance improves even further when using the recent attention-based Video-Swin backbone. Interestingly, even though our model does not use optical flow, it succeeds on this dataset where motion is the primary cue to segmentation due to the camouflaged nature of the animals. This fact supports the claim that our encoder is able to learn rich spatiotemporal features, even without flow input.

**DAVIS.** Table 2 shows DAVIS'16 results, with and without postprocessing. With the Video-Swin backbone, MED-VT outperforms all alternatives in mean/recall F-measure, $\mathcal{F}$, and mean/recall IoU, $\mathcal{J}$. When reverting to the ResNet101 backbone: Among approaches working directly on video frames (*i.e.* RGB without optical flow), MED-VT outperforms all alternatives on mean F-measure and mIoU. For the recall and decay measures we are comparable with all others. Additionally, our model without optical flow is on-par or even outperforms approaches that explicitly rely on optical flow, except for the recent RTNet [32]; although, even there we have a $1.1\%$ advantage on F-measure before postprocessing. Notably, our MED-VT relies only on RGB frames while other state-of-the-art approaches use optical flow as an additional input to exploit object motion. Moreover, most of the DAVIS'16 results include CRF postprocessing [16]. In contrast, we do not employ such complex postprocessing; rather, we follow a simpler multiscale inference strategy similar to that used by another approach [51].

There is evidence that success on DAVIS'16 is largely driven by the ability to capitalize on single frame/static appearance information (*e.g.* colour, texture), rather than dynamic (*e.g.* motion) information [15]. Unlike our MED-VT, RTNet [32], uses extra pre-training on a saliency segmentation dataset, which aligns with success on DAVIS'16 being tied to single frame information. Nevertheless, not only can MED-VT be competitive with RTNet (*e.g.* on mean F-measure without postprocessing) when using the same ResNet101 backbone, but by switching MED-VT to use the Video-Swin backbone we are able to yield better performance in boundary accuracy and also slightly better performance in mIoU, without an extra dataset or optical flow.

**YouTube-Objects**. Table 3 shows YouTube-Objects results. It is seen that our approach once again outper-

| Measures | | Uses RGB+Flow | | | Uses RGB only | | |
|---|---|---|---|---|---|---|---|
| | | COD (two-stream) [17] | MATNet [54] | RTNet [32] | COSNet [23] | **Ours** | **Ours†** |
| $\mathcal{J}$ Mean ↑ | | 55.3 | 64.2 | 60.7 | 50.7 | 69.4 | **77.9** |
| Success Rate ↑ | $\tau = 0.5$ | 0.602 | 0.712 | 0.679 | 0.588 | 0.762 | **0.874** |
| | $\tau = 0.6$ | 0.523 | 0.670 | 0.624 | 0.534 | 0.716 | **0.834** |
| | $\tau = 0.7$ | 0.413 | 0.599 | 0.536 | 0.457 | 0.657 | **0.777** |
| | $\tau = 0.8$ | 0.267 | 0.492 | 0.434 | 0.337 | 0.560 | **0.685** |
| | $\tau = 0.9$ | 0.088 | 0.246 | 0.239 | 0.167 | 0.369 | **0.440** |
| | $SR_{mean}$ | 0.379 | 0.544 | 0.502 | 0.417 | 0.613 | **0.722** |

Table 1. Results of moving camouflaged object segmentation on MoCA dataset with best overall results in **bold**. Results shown as mean Intersection over Union (mIoU) and localization success rate for various thresholds, $\tau$. Our results are reported with ResNet101 backbone, as used in the state of the art, as well as Video-Swin backbone, labelled with †.

| Measures | | Uses RGB + Optical Flow | | | | Uses RGB only | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EPO+ [1] | MATNet [54] | RTNet [32] | FSNet [11] | AGS [42] | COSNet [23] | AGNN [41] | ADNet [48] | DFNet [51] | **Ours** | **Ours†** |
| $\mathcal{J}$ | Mean ↑ | -/80.6 | -/82.4 | 84.3/85.6 | 82.1/83.4 | -/79.7 | -/80.5 | 78.9/80.7 | 78.26/81.7 | -/83.4 | 83.0/83.5 | 84.2/**85.9** |
| | Recall ↑ | -/95.2 | -/94.5 | -/96.1 | - | -/91.1 | -/94.0 | -/94.0 | - | - | 93.5/93.1 | 95.8/ **96.3** |
| | Decay ↓ | -/0.02 | -/5.5 | - | - | -/**0.0** | -/**0.0** | -/0.03 | - | - | 0.05/0.05 | 0.05/0.05 |
| $\mathcal{F}$ | Mean ↑ | -/75.5 | -/80.7 | 83.0/84.7 | 83.0/83.1 | -/77.4 | -/79.4 | -/79.1 | 77.1/80.5 | -/81.8 | 84.1/83.6 | 86.4/**86.6** |
| | Recall ↑ | -/87.9 | -/90.2 | -/93.8 | - | -/85.8 | -/90.4 | -/90.5 | - | - | 93.6/93.5 | 94.9/**95.1** |
| | Decay ↓ | -/0.02 | -/4.5 | - | - | -/**0.0** | -/**0.0** | -/0.03 | - | - | 0.03/0.03 | 0.03/ 0.03 |

Table 2. Results on DAVIS'16 validation set. For those using post processing (*e.g*. conditional random fields [1, 11, 23, 32, 41, 42, 51, 54], instance pruning [48, 51], multiscale inference [51]), results shown as $x/y$, with $x$ and $y$ results without and with post processing, resp. F-measure, $\mathcal{F}$, and mean Intersection over Union (mIoU), $\mathcal{J}$ are shown. We show our results with the standard ResNet-101 backbone and Video-Swin backbone, indicated by †. Best results highlighted in **bold**.

forms all others by a considerable margin when using the standard ResNet-101 backbone and further improves using Video-Swin. Per object category results are reported in the supplement, as standard for this dataset.

| Method | Input | mIoU |
|---|---|---|
| FSEG [10] | RGB + Optical Flow | 68.4 |
| LVO [37] | | 67.5 |
| MATNet [54] | | 69.0 |
| RTNet [32] | | 71.0 |
| PDB [35] | RGB | 65.4 |
| AGS [42] | | 69.7 |
| COSNet [23] | | 70.5 |
| AGNN [41] | | 70.8 |
| **MED-VT (Ours)** | RGB | 75.2 |
| **MED-VT (Ours†)** | | **78.5** |

Table 3. YouTube-Objects results given as mean Intersection over Union (mIoU); best results **bolded**. Our results shown with ResNet-101 backbone and Video-Swin backbone, indicated by †.

| Method | Input | Backbone | mIoU |
|---|---|---|---|
| Ji et al. [12] | RGB + Optical Flow | ResNet-101 | 36.9 |
| Dang et al. [7] | | ResNet-101 | 38.6 |
| SSA2D [31] | | I3D | 39.5 |
| **MED-VT (Ours)** | RGB | ResNet-101 | 39.5 |
| | | Video-Swin | **52.6** |

Table 4. State-of-the-art comparisons for actor/action segmentation on A2D dataset; best results in **bold**. Results given as mean Intersection over Union (mIoU).

| Decoder MS | Encoder MS | Label Propagation | DAVIS'16 | YouTube Objects | MoCA | A2D |
|---|---|---|---|---|---|---|
| - | - | - | 79.5 | 73.9 | 67.5 | 50.0 |
| ✓ | - | - | 81.5 | 74.2 | 67.7 | 50.9 |
| ✓ | ✓ | - | 82.2 | 74.4 | 69.1 | 51.6 |
| ✓ | ✓ | ✓ | **83.0** | **75.2** | **69.4** | **52.6** |

Table 5. Multiscale encoder-decoder and label propagation ablations reporting mIoU. Best results highlighted in **bold**.

**A2D: Actor/Action Segmentation**. Table 4 compares our approach on actor/action segmentation to a number of alternatives, which typically use optical flow as an extra input, across different feature backbones. Our results are consistently better or on par with the alternatives, even when we operate under the simplest setting, *i.e*. one input modality (RGB images) and weaker features, *i.e*. ResNet101 *vs*. I3D used by the previous state-of-the-art (SSA2D). When trained and tested with a stronger backbone (yet maintaining only RGB input), the improvements are even more notable, *e.g*., we outperform the previous best (SSA2D [31]) by more than 10% with Video-Swin features [22].

## 5.3. Ablation study

The previous section documented the overall strength of MED-VT, with an integral part of that presentation being a confirmation of our first contribution: state-of-the-art performance without extra optical flow input. In this section we conduct ablation experiments on both, AVOS (DAVIS'16) and actor/action segmentation (A2D), to investigate the two remaining contributions of our work: the multiscale encoder-decoder video transformer and many-to-

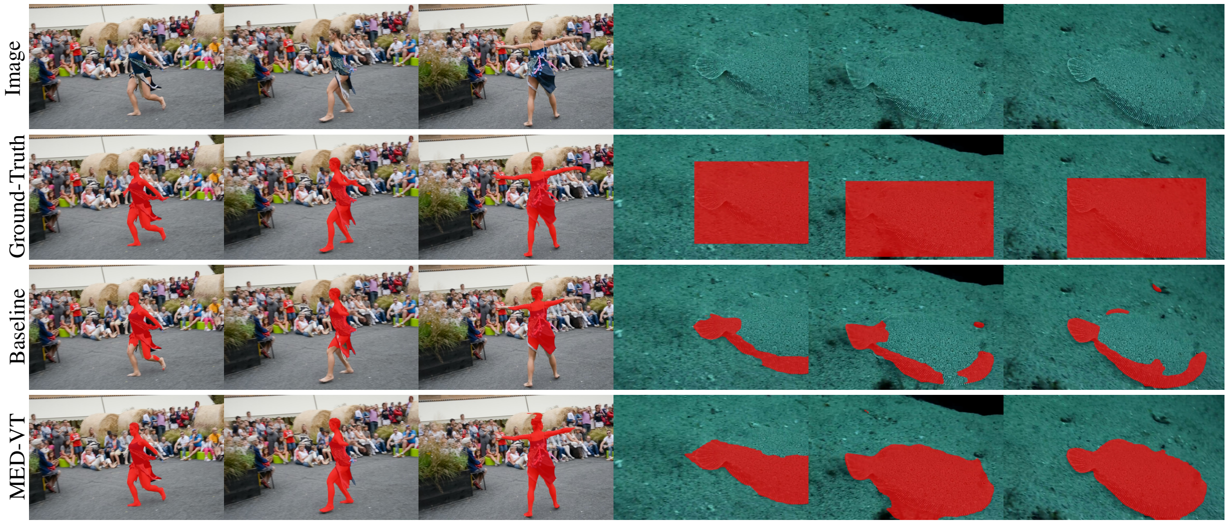

Figure 4. Qualitative segmentation results (red masks) comparing MED-VT to groundtruth and baseline algorithm. **Left:** Three frames of DAVIS'16 dance-twirl. **Right:** Three frames of MoCA flounder-6. MED-VT segments with fine precision and temporal consistency, even in the presence of severe camouflage. MoCA groundtruth only specified as bounding boxes; although, MED-VT goes further to delineate actual object shape. Supplemental video provides visualization especially useful for appreciating the camouflage example.

| Method | DAVIS'16 | YouTube Objects | MoCA |
|---|---|---|---|
| - | 82.2 | 74.4 | 69.1 |
| Mto1 | 81.7 | 74.5 | 68.6 |
| MtoM | **83.0** | **75.2** | **69.4** |

Table 6. Ablation on Many-to-One (Mto1) vs our Many-to-Many (MtoM) label propagation. Best results highlighted in **bold**.

many label propagation. To facilitate this study, we created a baseline model with a single scale transformer encoder and decoder working only on the top layer feature map of the backbone feature extractor, *i.e.* coarsest scale. Subsequently, we incrementally add the multiscale decoder, multiscale encoder and label propagation.

Table 5 shows that the multiscale decoder immediately improves over the baseline for both AVOS and actor/action segmentation. Addition of the multiscale encoder further improves the results to demonstrate their complementarity and the importance of unified multiscale encoding-decoding. Moreover, addition of many-to-many label propagation to the unified encoder-decoder consistently leads to the best overall performance. This enhancement can be traced to the label propagation yielding more temporally consistent predictions. Finally, we do a direct analysis of the benefits of our many-to-many label propagation vs. an alternative many-to-one label propagation technique, where many-to-one only propagates from previous frames to the current. Table 6 shows consistent improvement with many-to-many label propagation over the many-to-one alternative.

### 5.4. Qualitative results

Figure 4 shows qualitative results on two AVOS videos. The dancer on the left side exhibits complex, deforming motion and requires fine localization precision to delineate limbs. MED-VT deals with both challenges in a temporally consistent fashion, with consistency from the encoder as well as label propagator and localization from the decoder; in comparison, the baseline inconsistently captures the limbs, if at all. The fish on the right side is almost impossible to detect in a single frame due to its strong camouflage. MED-VT defeats the camouflage to precisely and consistently delineate the body as the encoder consistently abstracts critical motion, while the baseline largely fails; this example is best viewed in the supplemental video to reveal the input camouflaged fish. The supplemental video also has qualitative results for actor/action segmentation.

## 6. Conclusion

A novel Multiscale Encoder-Decoder Video Transformer (MED-VT) has been introduced. MED-VT is the first video transformer to unify multiscale encoding and decoding. Moreover, it is the first to apply many-to-many label propagation in a video transformer. The benefits of these innovations have been motivated and empirically validated. Encoding yields rich and temporally consistent spatiotemporal features derived from only RGB input (*i.e.* without optical flow). Decoding yields yields semantically informed, precise localization. Label propagation promotes consistency across an entire input clip. MED-VT has been instantiated on two video prediction tasks to yield state-of-the-art performance on multiple datasets. The generality of MED-VT also makes it suitable for other video-based dense prediction tasks where there may be little known a priori about the objects of interest, yet precise delineation is desired (*e.g.* anomalous behaviour detection in video [14, 27, 49] and multimodal video representation learning [18, 36]).

## References

[1] Ijaz Akhter, Mohsen Ali, Muhammad Faisal, and Richard Hartley. EpO-Net: Exploiting geometric constraints on dense trajectories for motion saliency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1273–1283, 2019. 7

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, pages 813–824. Proceedings of Machine Learning Research, 2021. 3

[3] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 19594–19607, 2021. 3

[4] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021. 1, 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 6

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1, 2, 4

[7] Kang Dang, Chunluan Zhou, Zhigang Tu, Michael Hoy, Justin Dauwels, and Junsong Yuan. Actor-action semantic segmentation with region masks. In *Proceedings of the British Machine Vision Conference*, 2018. 2, 7

[8] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 1, 2, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[10] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2126, 2017. 2, 7

[11] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4922–4933, 2021. 7

[12] Jingwei Ji, Shyamal Buch, Alvaro Soto, and Juan Carlos Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 702–717, 2018. 2, 7

[13] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–41, 2022. 1

[14] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental update. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009. 8

[15] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13999–14009, 2022. 6

[16] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, volume 24, pages 109–117, 2011. 5, 6

[17] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 488–503, 2020. 6, 7

[18] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 8

[19] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1, 2

[20] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2, 4

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2980–2988, 2017. 5

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 6, 7

[23] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3623–3632, 2019. 2, 7

[24] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2020. 2

[25] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9670–9679, 2021. 2, 5

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision*, pages 565–571, 2016. 5

[27] Trong Nguyen and Jean Meunier. Anomoly detection in video sequence with appearance motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1884–1893, 2019. 8

[28] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

[29] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 6

[30] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3289, 2012. 6

[31] Aayush J Rana and Yogesh S Rawat. We don't need thousand proposals: Single shot actor-action detection in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2960–2969, 2021. 7

[32] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15455–15464, 2021. 2, 6, 7

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2

[34] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (HRI) setting. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 50–56, 2019. 2

[35] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision*, pages 715–731, 2018. 7

[36] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 5, 8

[37] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4481–4490, 2017. 7

[38] Vladimir Vapnik. Transductive inference and semi-supervised learning. In *Semi-Supervised Learning*, chapter 24, page 454–472. MIT press, 2006. 1

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, volume 30, 2017. 5

[40] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2019. 2

[41] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9236–9245, 2019. 2, 7

[42] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3074, 2019. 7

[43] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *International Conference on Learning Representations*, 2022. 1, 2

[44] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 3

[45] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021. 2

[46] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015. 2

[47] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2264–2273, 2015. 6

[48] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the*

*IEEE/CVF International Conference on Computer Vision*, pages 931–940, 2019. 7

[49] Dan Yu, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 8

[50] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2020. 2

[51] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. Learning discriminative feature with CRF for unsupervised video object segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 445–462, 2020. 6, 7

[52] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, volume 16, 2003. 1, 5

[53] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[54] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13066–13073, 2020. 2, 6, 7

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 2