Temporal-aware Hierarchical Mask Classification for Video Semantic Segmentation

Zhaochong An[1]
zhaoan@ethz.ch

Guolei Sun[1*]
guolei.sun@vision.ee.ethz.ch

Zongwei Wu[2]
zongwei.wu@uni-wuerzburg.de

Hao Tang[1]
hao.tang@vision.ee.ethz.ch

Luc Van Gool[1,3]
vangool@vision.ee.ethz.ch

[1] Computer Vision Lab,
ETH Zurich,
Zurich, Switzerland

[2] Computer Vision Lab, CAIDAS & IFI,
University of Wurzburg,
Wurzburg, Germany

[3] VISICS,
KU Leuven,
Leuven, Belgium

Abstract

Modern approaches have proved the huge potential of addressing semantic segmentation as a mask classification task which is widely used in instance-level segmentation. This paradigm trains models by assigning part of object queries to ground truths via conventional one-to-one matching. However, we observe that the popular video semantic segmentation (VSS) dataset has limited categories per video, meaning less than 10% of queries could be matched to receive meaningful gradient updates during VSS training. This inefficiency limits the full expressive potential of all queries. Thus, we present a novel solution **THE-Mask** for VSS, which introduces **t**emporal-aware **hie**rarchical object queries for the first time. Specifically, we propose to use a simple **two-round matching mechanism** to involve more queries matched with minimal cost during training while without any extra cost during inference. To support our more-to-one assignment, in terms of the matching results, we further design a **hierarchical loss** to train queries with their corresponding hierarchy of primary or secondary. Moreover, to effectively capture temporal information across frames, we propose a **temporal aggregation decoder** that fits seamlessly into the mask-classification paradigm for VSS. Utilizing temporal-sensitive multi-level queries, our method achieves state-of-the-art performance on the latest challenging VSS benchmark VSPW without bells and whistles. The code is available at github.com/ZhaochongAn/THE-Mask.

1 Introduction

Video semantic segmentation (VSS) is to assign per-pixel semantic categories to each frame of a video. As a fundamental task of scene understanding, VSS has significant implications
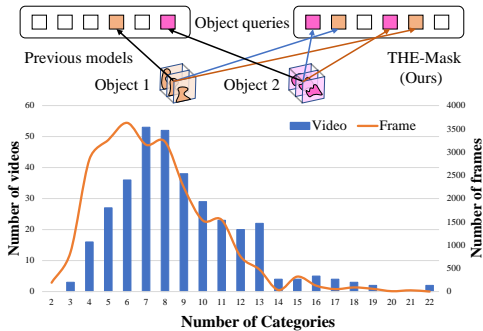

Figure 1: In VSPW, each frame/video has about 8 categories on average. This means only 8 out of 100 queries in previous mask-classification-based models (*e.g.*, Mask2Former [13]) can be matched to train effectively in one iteration, which limits fully utilizing the expression ability of all queries. By contrast, our THE-Mask can involve $2\times$ matched queries with rich semantic hierarchy for better performance.

for wide applications such as image editing [25], autonomous driving [20], and medical diagnosing [83]. Besides, the recent release of the large-scale dataset VSPW [52], with its higher annotated frame rate, has further spurred advancements in VSS.

VSS can be seen as extending image semantic segmentation to the video domain. Unlike images, videos contain crucial temporal semantic information, which helps segment objects despite their motion blur and occlusions across frames. Therefore, simply applying image-level segmentation models [29, 59] to videos yields suboptimal performance. Several methods [40, 57, 79] adapt frame-wise predictions for high temporal consistency, while others use per-clip approaches [21, 52, 62, 72] trained with video clips to aggregate temporal features. Despite these efforts, VSS has been mainly treated as a per-pixel classification task. However, recent works [12, 13, 41] show that mask-classification architecture, commonly used in instance segmentation, can also achieve satisfactory results for semantic segmentation. For example, Mask2Former [13] follows DETR [4] to learn object-centric representations as queries that are transformed into final image segments. These representations [4, 47] have shown their effectiveness in capturing object features for image input. The insight inspires us to further explore the potential of mask-based approaches for the VSS task.

In this work, we rethink the application of the mask classification paradigm in VSS. This paradigm predicts for each learned object query a pair of a binary mask and a class distribution. The final segmentations are aggregated from all predicted pairs. By default, the number of object queries is usually set to be a large number (*e.g.*, 100 in [13]). During training, most existing methods [4, 12, 13, 41] adopt a one-to-one bipartite matching, *i.e.*, Hungarian matching [38], to assign the best-fit predicted pair to each ground truth pair. Regarding the loss, the matched predicted pairs are optimized with the ground truths, while the unmatched pairs are forced to predict the artificial "no-object" category. In such a manner, only matched queries can be trained effectively while unmatched queries receive meaningless updates. However, as shown in Fig. 1, in the benchmark VSPW dataset, only about 8% queries can be matched on average to receive informative gradient updates for each training sample. This hinders fully utilizing the representation ability of all queries and harms performance. Given this observation, one question emerges: *is there a simple method to involve more queries during training without losing their own object representation abilities?* Besides, to leverage the temporal consistency in the frame domain, the second question naturally raises: *How to effectively model the temporal interactions in the mask-classification-based paradigm?*

To address the above questions, we propose a novel framework, termed **THE-Mask**, for VSS. For the first question, we introduce a **two-round matching mechanism** to bind more queries with ground truths, yielding a more-to-one matching manner for VSS. To support this paradigm, we propose a **hierarchical loss** which enables each query to learn its specific

non-overlapping object representation. This increases the expression variety of matched object queries and results in hierarchical queries with rich semantics at different levels. These complementary queries contribute to the final segments from both primary and secondary views, leading to improved performance. For the second question, we propose a lightweight **Temporal Aggregation Decoder (TAD)** that utilizes two sets of queries, *i.e.*, video-level and image-level, to explicitly model image-video relationships and learn temporal information, optimizing the interactions between queries and multiple frames. TAD effectively avoids hard long attention sequences while adding only 3.2M parameters over the baseline counterpart.

We evaluate THE-Mask against the state-of-the-art methods on the most challenging VSS dataset VSPW [52]. Our THE-Mask trained on 4-frame clips ($t = 4$) achieves the state-of-the-art result of 49.1% mIoU (MiT-B2 [71] as Backbone), surpassing the prior method MRCFA [53] by 3.8% mIoU. Compared to our baseline [13], our method achieves a 0.9% absolute gain in mIoU ($t = 1$) without any extra inference cost and further achieves better performance with longer clips. We also perform a detailed ablation study to validate the effectiveness of our approach. Our experiments demonstrate the efficacy of THE-Mask and highlight the potential of mask-classification models for the VSS task. We believe that our model is an effective baseline for future VSS research. In summary, the contributions of THE-Mask are as follows:

- Our THE-Mask is the first of its kind to introduce hierarchy into object queries in the mask-classification-based paradigm. The parameter-free hierarchical design enriches the expression ability of queries and improves the performance without any extra inference cost.

- To leverage the temporal clues in the video setting, we propose a temporal aggregation decoder to effectively model cross-frame interactions while fitting seamlessly into the mask-classification paradigm. THE-Mask achieves the new state-of-the-art performance on the VSPW benchmark.

2 Related Work

Image semantic segmentation (ISS) is to assign a semantic label to each pixel of the input image. Naturally, it can be formulated as a per-pixel classification task. From the early FCNs [48], most works follow the per-pixel setting and differ to exploit semantic information in proposing new structures to enlarge the receptive field [6, 15, 56, 75]; designing multi-scale feature ensemble methods [7, 8, 9, 10, 22, 23, 43, 45, 58, 69, 74, 80]; aggregating stored dataset-level representations [53]; using non-local context aggregation schemes [18, 27, 54, 61, 77, 81]; or utilizing long-range modeling capacity of transformers [60, 71, 82]. More recently, motivated by DETR [4], MaskFormer [12] and Mask2Former [13] address ISS using query-based transformer architectures with a mask-classification-based paradigm, which has been widely used in instance-level segmentation [3, 5, 11, 24, 64, 66, 67, 68]. The success of the mask-classification-based perspective for ISS inspires us to explore the new paradigm in the video domain.

Video semantic segmentation (VSS) as an extension of ISS aims to predict pixel-level semantics in consecutive video frames. Some works treat VSS in a per-frame fashion and refine the predictions for temporal consistency [29, 30, 39, 40, 57, 79]. Other works explore different mechanisms in training to fuse semantic information across time by using patch
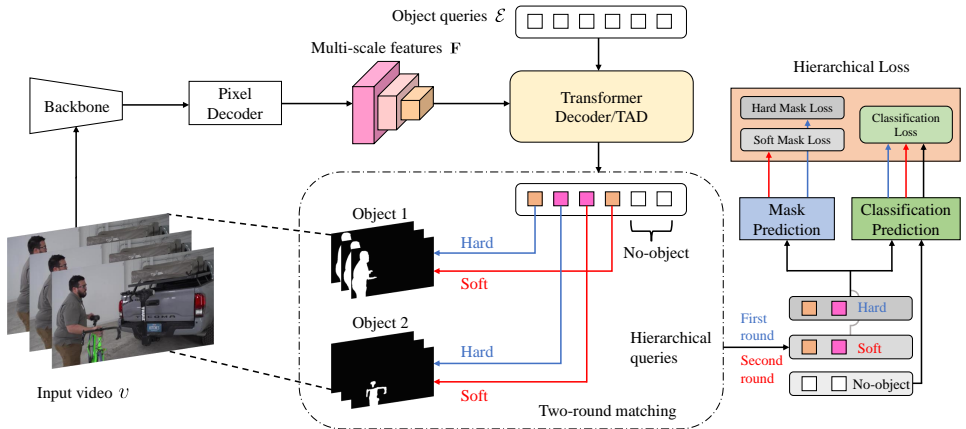
Figure 2: **THE-Mask** uses the mask-classification-based architecture with a backbone, a pixel decoder, and a transformer decoder. To fully exploit the expression ability of queries, we propose a hierarchical loss, which can effectively train $2\times$ queries as before in one training iteration with richer semantic features. Beside, to learn the temporal clues, we design the Temporal Aggregation Decoder (TAD) to replace the original transformer decoder.

matching or optical flow for label propagation [1, 2, 54]; utilizing the predictive learning features carrying the temporal context [32, 50]; employing recurrent units to propagate features from past frames to current frame [17, 55]; exploiting aligned previous segmentation maps as supervised signals [21, 46, 72, 85]; aggregating the dataset-level representations of previous frames [85]; or warping representations of adjacent frames [16, 19, 42, 44, 52, 62, 63] by using various temporal-adapted wrapping modules [26, 76, 80]. Beyond exploring more temporal-accurate models, another line of research focuses on improving the efficiency [28, 31, 46, 51, 73, 84]. Recently, a few works have applied query-styled approaches in the video domain as well. [36] built upon mask transformers [66] introduces a latent memory to facilitate attention learning over multiple frames. Video k-net [41] adapts K-net [28] into the video by performing kernel interactions along the temporal dimensions.

3 Proposed Method

3.1 Preliminaries

Mask classification formulation. Given an input video $v \in \mathbb{R}^{T \times 3 \times H \times W}$ with T frames of spatial size $H \times W$, the traditional VSS models predict the probability distribution over all categories $\mathbb{C} = \{1, ..., K\}$ for each video pixel: $\{p_i | p_i \in \Delta^K\}_{i=1}^{T \cdot H \cdot W}$. Here, Δ^K is the K-dimensional probability simplex. Different from the above per-pixel classification manner, we use the mask classification paradigm, which learns a set of N queries $\mathcal{E} = \{q_1, ..., q_N | q_i \in \mathbb{R}^C\}$ as object-centric representations to output N classification-mask pairs:

$$\{\hat{y}_i\}_{i=1}^N = \{(\hat{p}_i, \hat{m}_i) | \hat{m}_i \in \{0, 1\}^{T \times H \times W}, \quad \hat{p}_i \in \Delta^{K+1}\}_{i=1}^N, \tag{1}$$

where $\hat{m}_i$ is the predicted mask from object query q_i, and $\hat{p}_i$ containing K categories and a "no object" label ($\varnothing$) is the class probability distribution assigned to $\hat{m}_i$. Let us denote $\hat{p}_i(c)$ as the probability of assigning class $c \in \mathbb{C}$ to mask $\hat{m}_i$. During inference [12, 13], it will

aggregate all the predicted pairs $\{\hat{y}_i\}_{i=1}^N$ to the segmentation output $\hat{y}_{t,h,w}$ for pixel (h,w) at frame t by: $\hat{y}_{t,h,w} = \arg\max_{c \in \mathbb{C}} \sum_{i=1}^N \hat{p}_i(c) \cdot \hat{m}_{i,t,h,w}$.

Mask classification models firstly use a *backbone* to extract the features of each input frame individually, then refine the features via a *pixel decoder*, and finally apply a *transformer decoder* to learn the object queries and output $\{\hat{y}_i\}_{i=1}^N$ in parallel. Training such models need a one-to-one bipartite matching $\sigma = \{\sigma(i)\}_{i=1}^{N^{gt}}$ [4, 58, 59] to assign the best-fit N^{gt} predictions from $\{\hat{y}_i\}_{i=1}^N$ to the ground truth set $\{y_i = (c_i, m_i)\}_{i=1}^{N^{gt}}$, where $c_i^{gt} \in \mathbb{C}$ is the ground truth class label of mask $m_i \in \{0,1\}^{T \times H \times W}$. Thus, we denote $\hat{y}_{\sigma(i)}$ as the matched pair to y_i. The matching score $\mathcal{S}_{match}$ used for evaluating the fitness between $\hat{y}_j$ and y_i is defined as:

$$\mathcal{S}_{match}(\hat{y}_j, y_i) = \underbrace{\lambda_{ce}\mathcal{L}_{ce}(\hat{m}_j, m_i) + \lambda_{dice}\mathcal{L}_{dice}(\hat{m}_j, m_i)}_{\mathcal{L}_{mask}} - \hat{p}_j(c_i), \tag{2}$$

where $\mathcal{L}_{ce}$ is the binary cross-entropy loss and $\mathcal{L}_{dice}$ is the dice loss [53] with weights λ_{ce}, and λ_{dice}. The remaining $N - N^{gt}$ queries are matched to $\varnothing$ class. Then the optimization of model parameters θ is via minimizing the final loss $\mathcal{L}$ over all queries, consisting of $\mathcal{L}_{mask}^{\sigma}$ over matched queries in σ and a classification term $\mathcal{L}_{cls}$ for unmatched queries to predict $\varnothing$ and for matched queries to ground truth labels:

$$\min_{\theta} \mathcal{L} = \sum_{i=1}^{N^{gt}} \mathcal{L}_{mask}^{\sigma}(\hat{m}_{\sigma(i)}, m_i) + \underbrace{(-\sum_{i=1}^{N^{gt}} \log \hat{p}_{\sigma(i)}(c_i) - \sum_{i \notin \sigma} \log \hat{p}_i(\varnothing))}_{\mathcal{L}_{cls}}. \tag{3}$$

3.2 Hierarchical Mask Classification

To involve more queries into training, we design an effective two-round matching with hierarchical loss functions to leverage richer expression ability (Fig. 2).

Two-round queries. In order to give meaningful gradient updates to more queries during training, we propose to use two-round matching. After getting the first round matching σ_1 according to Eq. (2) through Hungarian algorithm [58], we do a second matching among the remaining queries to get σ_2 where we only use $\mathcal{L}_{ce}$ as the matching score. We simply assign the unmatched queries after two rounds to predict $\varnothing$. Thus, we easily include 2× matched queries as many as before into training by supervising them with informative ground truths. This design is time-efficient since we can get the second matching results directly by reusing the cost matrix $\mathcal{L}_{ce}$ from the first matching round.

Hierarchical loss functions. For σ_1 and σ_2, if we use the same mask loss $\mathcal{L}_{mask}$ on them, it will make the learned semantic features of $q_{\sigma_1(i)}$ and $q_{\sigma_2(i)}$ overlap. Then the total semantic variety of queries still remains limited. So, we further design the hierarchical loss with different emphasizes on the two matched groups. For σ_1 which contains the best-fit queries towards ground truths, we expect queries to take primary responsibility to segment the whole objects. So, we use the **hard** mask classification loss $\mathcal{L}_{hard}$ for queries in σ_1:

$$\mathcal{L}_{hard} = \sum_{i=1}^{N^{gt}} [-\log \hat{p}_{\sigma_1(i)}(c_i) + \mathcal{L}_{mask}^{\sigma_1}(\hat{m}_{\sigma_1(i)}, m_i)], \tag{4}$$

where $\mathcal{L}_{mask}^{\sigma_1}$ includes the weighting hyper-parameters $\lambda_{ce}^{\sigma_1}$ and $\lambda_{dice}^{\sigma_1}$ as in Eq. (2).

For σ_2, each query as the second best-fit can also represent the matched object partially. Not required to learn all the semantic features of the object, instead we design the **soft** mask

classification loss $\mathcal{L}_{soft}$ to help it refine its currently learned knowledge about that object. $\mathcal{L}_{soft}$ has the similar formula as $\mathcal{L}_{hard}$ except using soft mask sm_i to replace m_i in Eq. (4):

$$\mathcal{L}_{soft} = \sum_{i=1}^{N^{gt}}[-\log \hat{p}_{\sigma_2(i)}(c_i) + \mathcal{L}_{mask}^{\sigma_2}(\hat{m}_{\sigma_2(i)}, sm_i)], \text{where } sm_i = \hat{m}_{\sigma_2(i)} \circ m_i, \quad (5)$$

where $\circ$ is the Hadamard product. The soft mask is generated from the intersection area between the matched ground truth mask and the predicted mask, which pushes the query to focus on consolidating its learned semantic features on the matched object.

In our final hierarchical loss $\mathcal{L}_{hie}$ used by THE-Mask, we balance the two mask terms in $\mathcal{L}_{hard}$ and $\mathcal{L}_{soft}$ by a round weight α and merge all the classification terms from three query groups, *i.e.*, hard group, soft group, and unmatched group, into one classification term $\mathcal{L}_{cls}$:

$$\mathcal{L}_{hie} = \sum_{i=1}^{N^{gt}}[\mathcal{L}_{mask}^{\sigma_1}(\hat{m}_{\sigma_1(i)}, m_i) + \alpha\mathcal{L}_{mask}^{\sigma_2}(\hat{m}_{\sigma_2(i)}, sm_i)] + \mathcal{L}_{cls}. \quad (6)$$

Our hierarchical loss introduces the hierarchical structures in queries by considering different semantic requirements in terms of the fitness of object queries to ground truths and optimizing the matching relationship between queries and objects, which is critical for richer expression ability and higher accuracy.
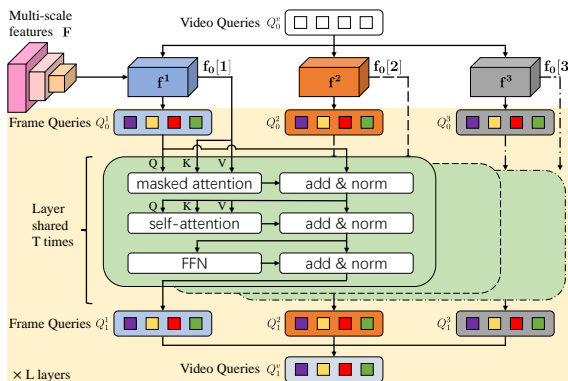
3.3 Temporal Aggregation Decoder

Figure 3: The architecture of TAD.

Since VSS requires detecting objects across frames, it is vital to learn the semantic information from multiple frames jointly for temporal consistency. Denote the multi-scale features from pixel decoder as $\mathbf{F} = \{\mathbf{f_i}|\mathbf{f_i} \in \mathbb{R}^{T \times H_i \times W_i \times C}\}_{i=1}^n$, where $\mathbf{f_i}[t] \in \mathbb{R}^{H_i \times W_i \times C}$ indicates the i-th scale features of t-frame. One naive way to model the temporal context is to directly attend queries to the features of T frames $\mathbf{f_i}$. But the T frames token sequence is very long and thus it is very hard to learn the temporal information.

To alleviate the issue, we propose the lightweight **Temporal Aggregation Decoder** (**TAD**) as in Fig. 3 to effectively model the temporal interactions while doing the attention at the frame level. In this decoder, we initiate one set of video-level queries $Q_0^v = \{q_{0,i}^v\}_{i=1}^N \in \mathbb{R}^{N \times C}$ which is responsible for video-level features. We flatten and concatenate the multi-scale features $\{\mathbf{f_i}\}_{i=1}^n$ to form frame features $\{\mathbf{f^t}|\mathbf{f^t} \in \mathbb{R}^{(\sum_{i=1}^n H_i W_i) \times C}\}_{t=1}^T$. Then before passing into the decoder layer, Q_0^v will attend to the frame features $\mathbf{f^t}$ to produce the frame-level queries $Q_0^t = \{q_{0,i}^t\}_{i=1}^N \in \mathbb{R}^{N \times C}$ which are only responsible to interact with the t-th frame:

$$Q_0^t = \text{SM}(\frac{f_Q(Q_0^v)f_K(\mathbf{f^t})^T}{\sqrt{C}})f_V(\mathbf{f^t}), \quad (7)$$

where SM means Softmax, f_Q, f_K, and f_V are linear maps to generate query, key, and value respectively. For simplicity, we omit the formulation of multi-head attention [55] here.

Then we pass one resolution of the multi-scale features $\mathbf{f_l}$ into l-th decoder layer at a time in a round-robin fashion [13]. In each layer, we do the masked attention at the frame level:

$$Q_l^t = Q_{l-1}^t + \text{SM}(\mathcal{M}_{l-1}^t + \frac{f_Q(Q_{l-1}^t)f_K(\mathbf{f_l[t]})^T}{\sqrt{C}})f_V(\mathbf{f_l[t]}), \quad \mathcal{M}_{l-1}^t = \begin{cases} 0 & \text{if } M_{l-1}^t(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{8}$$

where Q_l^t denotes the frame-level queries at l-th layer, and $M_{l-1}^t \in \{0,1\}^{N \times H_l W_l}$ is the mask prediction from Q_{l-1}^t resized to the resolution of $\mathbf{f_l[t]}$. After this frame-level attention of each layer, we aggregate all frame-level queries $q_{l,i}^t$ to our video-level queries $q_{l,i}^v$:

$$q_{l,i}^v = q_{l-1,i}^v + \frac{\sum_{t=1}^T q_{l,i}^t \times \exp(\text{FC}(q_{l,i}^t))}{\sum_{t=1}^T \exp(\text{FC}(q_{l,i}^t))}, \tag{9}$$

where FC is a fully connected layer to reduce channels to 1 and a softmax is used to obtain the weight on each $q_{l,i}^t$ over T frames. The aggregation operation enables video-level queries to collect frame-level information and learn the globally temporal-sensitive representations.

During training, we combine the class predictions from video-level queries and mask predictions from frame-level queries as input $\{\hat{y}_i = (\hat{p}_i^v, \{\hat{m}_i^t\}_{t=1}^T)\}_{i=1}^N$ (ignore l for simplicity) to the two-round matching module to get both groups involved in the matching phase. For inference, we simply use the average of predictions from video-level and frame-level queries:

$$\hat{y}_{t,h,w} = \arg\max_{c \in \mathbb{C}} \sum_{i=1}^N \frac{1}{2}(\hat{p}_i^v(c) \cdot \hat{m}_{i,t,h,w}^v + \hat{p}_i^t(c) \cdot \hat{m}_{i,h,w}^t). \tag{10}$$

In this way, we could avoid the long attention sequence and model the temporal information through the interactions between video-level and frame-level queries effectively.

4 Experiments

4.1 Implementation Details

THE-Mask is implemented using MMSegmentation toolbox [14]. Following [52], we use the encoder of SegFormer [71] as our backbones, which is a hierarchical transformer pretrained on ImageNet [37]. The pixel decoder and transformer decoder follow Mask2Former [13].

For our hierarchical loss, we set $\lambda_{ce}^{\sigma_1} = \lambda_{dice}^{\sigma_1} = 2.5, \lambda_{ce}^{\sigma_2} = \lambda_{dice}^{\sigma_2} = 0.25$ and $\alpha = 0.5$. We use AdamW [49] and the *poly* [8] learning rate schedule with an initial learning rate of 10^{-4} and a weight decay of 0.05. A learning rate multiplier of 0.1 is applied to our backbone. We adopt data augmentations including standard random scale jittering with a resizing scale sampled from 0.5 to 2.0 followed by random cropping to 480×480, random horizontal flipping with probability 0.5, and standard random color jittering. We calculate $\mathcal{L}_{mask}$ with sampled points in both matching scores and final losses following [13]. We train models using 4 NVIDIA GPUs with a batch size of 8 (2 clips per GPU). For 1-frame clips training ($t = 1$), we use the first 30k as warm-up iterations with the original one-round matching and loss. After 30k, we apply our hierarchical loss for 130k iterations. For multiple-frame clips training ($t > 1$), we replace the original transformer decoder with TAD, load the weights

Method	Backbone	Params (M) ↓	mIoU ↑	Weighted IoU ↑
DeepLabv3+ [9]	ResNet-101	62.7	34.7	58.8
UperNet [70]	ResNet-101	83.2	36.5	58.6
PSPNet [80]	ResNet-101	70.5	36.5	58.1
OCRNet [76]	ResNet-101	58.1	36.7	59.2
ETC [46]	PSPNet	89.4	36.6	58.3
ETC [46]	OCRNet	58.1	37.5	59.1
NetWarp [70]	PSPNet	89.4	37.0	57.9
NetWarp [70]	OCRNet	58.1	37.5	58.9
TCB$_{st-ppm}$ [52]	ResNet-101	70.5	37.5	58.6
TCB$_{st-ocr}$ [52]	ResNet-101	58.1	37.4	59.3
TCB$_{st-ocr-mem}$ [52]	ResNet-101	58.1	37.8	59.5
Video K-Net (Deeplabv3+) [41]	ResNet-101	_	37.9	_
Video K-Net (PSPNet) [41]	ResNet-101	_	38.0	_
SegFormer [71]	MiT-B1	13.8	36.5	58.8
SegFormer [71]	MiT-B2	24.8	43.9	63.7
CFFM ($t = 4$) [72]	MiT-B1	15.5	38.5	60.0
CFFM ($t = 4$) [72]	MiT-B2	26.5	44.9	64.9
MRCFA ($t = 4$) [63]	MiT-B1	16.2	38.9	60.0
MRCFA ($t = 4$) [63]	MiT-B2	27.3	45.3	64.7
	MiT-B0	23.0	38.9	60.9
Mask2Fomer ($t = 1$) [13]	MiT-B1	33.0	43.3	63.6
	MiT-B2	44.0	47.6	65.4
	MiT-B0	23.0	**39.8**	**61.3**
THE-Mask ($t = 1$)	MiT-B1	33.0	**44.1**	**64.2**
	MiT-B2	44.0	**48.5**	**66.2**
THE-Mask ($t = 2$)	MiT-B5	104.5	**52.1**	**67.2**

Table 1: **Comparison with state-of-the-art methods on the VSPW validation set.** Our model outperforms both the best per-pixel classification approaches and the strong mask classification-based baseline.

backbone	$t = 1$	$t = 2$	$t = 4$
MiT-B0	39.76	39.94	**40.68**
MiT-B1	44.06	44.68	**45.19**
MiT-B2	48.53	48.99	**49.11**

Table 2: Effects of training clip length.

temporal setting	$t = 2$	$t = 4$
one-to-video	43.40	43.90
one-to-frame	44.41	44.84
video-frame	**44.68**	**45.19**

Table 3: Temporal aggregation ablation.

from the $t = 1$ model, and finetune TAD by freezing the backbone and pixel decoder. The finetune setting is the same as above except using only 12K iterations which are very fast. For testing, we resize all frames on VSPW to 480×853 and conduct single-scale inference. THE-Mask is flexible to infer a video of arbitrary length without any post-processing. By default, we divide one video into non-overlapping clips as long as the training clips. For results, we report the mean of three runs.

4.2 Main Results

The comparison of THE-Mask with state-of-the-art methods on VSPW [52] are listed in Table 1. From DeepLabv3+ [9] to MRCFA [63], they are all per-pixel classification paradigms. We select the mask-classification-based Mask2Fomer [13] as the baseline model. From Table 1, we can observe that the mask-classification-based models perform much better than

previous per-pixel classification paradigms. The baseline model Mask2Fomer trained on 1-frame clips outperforms MRCFA trained on 4-frame clips with backbone MiT-B1 and MiT-B2 by 4.4% and 2.3% mIoU, respectively. It shows the huge potential of the mask-classification-based model over the per-pixel classification paradigm for the VSS task.

In the comparison between THE-Mask ($t = 1$) and the baseline, THE-Mask ($t = 1$) consistently boosts Mask2Fomer by a significant margin on both mIoU (0.8-0.9%) and wIoU (0.4-0.8%) without any extra-processing steps while having the same amount of parameters. Note that THE-Mask ($t = 1$) does not use the TAD since the one-frame training clips do not provide temporal information. So, the only difference between THE-Mask ($t = 1$) and Mask2Fomer lies in the hierarchical query design. It shows that our parameter-free hierarchical design can better utilize object queries for learning richer semantic features and achieve superior performance effectively. We further use a larger backbone (MiT-B5) in THE-Mask ($t = 2$) with TAD trained on two-frame clips to learn temporal clues. It sets the new state-of-the-art results of 52.1 mIoU and 67.2 wIoU.

4.3 Ablation Studies

Here, we provide ablation studies for more insights into the effects of different designs. The experiments are performed on VSPW [52] validation set using a MiT-B1 backbone.

Matching	Loss	mIoU ↑	wIoU ↑
one round	original loss	43.26	63.56
two round	original loss	43.70	63.77
two round	hierarchical loss	**44.06**	**64.16**

Table 4: Ablation study on hierarchical loss.

Training clip length. Table 2 (mIoU reported) shows how the length of training clips affects the performance of THE-Mask. When training on multiple-frame clips, we use the TAD to mine temporal information. On all three backbones, TAD clearly brings us better results when using longer training clips that contain richer semantic context to help segmentation. It demonstrates the effectiveness of the proposed TAD to learn temporal clues from long clips.

Temporal aggregation. In Table 3 (mIoU reported), we present ablation on different temporal learning settings of THE-Mask. Here, one-to-video means one set of object queries attends to the features of all frames. One-to-frame represents attending one set of queries to each frame by replicating that query set. And video-frame means TAD where we use the interaction between two sets of queries, *i.e.*, video-level and frame-level queries, to aggregate temporal information. One-to-frame yields 1.11% ($t = 2$) and 0.94% ($t = 4$) mIoU improvements than one-to-video. It shows that directly attending queries to all frames is difficult to learn temporal clues due to the long token sequence of multiple frames. Change to video-frame setting with only 3.2M extra parameters involved further improves mIoU by 0.27% ($t = 2$) and 0.35% ($t = 4$), suggesting the separation of attention and explicitly modeling of frame aggregation are key factors for temporal learning.

Hierarchical queries. In Table 4, we ablate our hierarchical query design to verify its effectiveness. The models are trained by 1-frame clips. We develop a variant by adding second-round matching and using the same original (hard) loss to the second-matched group. As mentioned earlier, simply adding one more matching round could involve more queries in the training, but the same loss applied to the two matched groups still limits the full utilization of all queries. When we apply our proposed hierarchical loss, it brings a large improvement of 0.8% mIoU and 0.6% wIoU to the baseline model. It shows that our proposed parameter-free hierarchical design can effectively supervise the two query groups with different focuses, and the hierarchy learned inside the object queries is helpful for performance gains.

5   Conclusion

In this paper, we rethink the application of mask-classification-based models in the VSS domain. Based on the fact of low utilization of object queries during training, we present THE-Mask, a simple and strong mask-classification-based model for VSS. THE-Mask is the first model to renovate the traditional one-to-one matching and introduce hierarchical structures into queries to fully exploit the representation ability of queries. Besides, we propose TAD to explicitly model the temporal interactions for cross-frame learning. Equipped with the parameter-free hierarchical design optimizing the matching relationship between queries and objects, and TAD optimizing the interactions between queries and multiple frames, THE-Mask achieves the state-of-the-art results on the VSPW dataset. We hope our methods can inspire future research in the segmentation domain.

References

[1] Vijay Badrinarayanan, Fabio Galasso, and Roberto Cipolla. Label propagation in video sequences. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3265–3272. IEEE, 2010.

[2] Ignas Budvytis, Patrick Sauer, Thomas Roddick, Kesar Breen, and Roberto Cipolla. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 230–237, 2017.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.7062.

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[8] 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.

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[11] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[12] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[13] 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.

[14] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. *Availabe online: https://github. com/openmmlab/mmsegmentation (accessed on 18 May 2022)*, 2020.

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[16] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every frame counts: Joint learning of video segmentation and optical flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10713–10720, 2020.

[17] Mohsen Fayyaz, Mohammad Hajizadeh Saffar, Mohammad Sabokrou, Mahmood Fathy, Fay Huang, and Reinhard Klette. Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes. In *Asian Conference on Computer Vision*, pages 493–509. Springer, 2016.

[18] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[19] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4462, 2017.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[21] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8053–8064, 2021.

[22] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[23] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3562–3572, 2019.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[25] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[27] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[28] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020.

[29] Po-Yu Huang, Wan-Ting Hsu, Chun-Yueh Chiu, Ting-Fan Wu, and Min Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[30] Junhwa Hur and Stefan Roth. Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*, pages 163–177. Springer, 2016.

[31] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019.

[32] Xiaojie Jin, Xin Li, Huaxin Xiao, Xiaohui Shen, Zhe Lin, Jimei Yang, Yunpeng Chen, Jian Dong, Luoqi Liu, Zequn Jie, et al. Video scene parsing with predictive feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5580–5588, 2017.

[33] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7231–7241, 2021.

[34] Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7189–7198, 2021.

[35] Zhenchao Jin, Dongdong Yu, Zehuan Yuan, and Lequan Yu. Mcibi++: Soft mining contextual information beyond image for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[36] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13914–13924, 2022.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[38] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[39] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3168–3175, 2016.

[40] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020.

[41] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857, 2022.

[42] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018.

[43] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[44] Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. Surveillance video parsing with single frame supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–421, 2017.

[45] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[46] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pages 352–368. Springer, 2020.

[47] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33: 11525–11538, 2020.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[50] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 648–657, 2017.

[51] Behrooz Mahasseni, Sinisa Todorovic, and Alan Fern. Budget-aware deep semantic video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2017.

[52] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021.

[53] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[54] Siva Karthik Mustikovela, Michael Ying Yang, and Carsten Rother. Can ground truth label propagation from video help semantic segmentation? In *European Conference on Computer Vision*, pages 804–820. Springer, 2016.

[55] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018.

[56] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[57] Hyojin Park, Alan Yessenbayev, Tushar Singhal, Navin Kumar Adhikari, Yizhe Zhang, Shubhankar Mangesh Borse, Hong Cai, Nilesh Prasad Pandey, Fei Yin, Frank Mayer, et al. Real-time, accurate, and consistent video semantic segmentation via unsupervised adaptation and cross-unit deployment on mobile device. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21431–21438, 2022.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[59] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[60] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[61] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European conference on computer vision*, pages 347–365. Springer, 2020.

[62] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3126–3137, 2022.

[63] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *European Conference on Computer Vision*, pages 522–539. Springer, 2022.

[64] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020.

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[66] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021.

[67] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

[68] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *CoRR*, abs/2112.08275, 2021. URL https://arxiv.org/abs/2112.08275.

[69] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *European Conference on Computer Vision*, pages 648–663. Springer, 2016.

[70] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[71] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[72] Yun Xing, Dayan Guan, Jiaxing Huang, and Shijian Lu. Domain adaptive video segmentation via temporal pseudo supervision. *arXiv preprint arXiv:2207.02372*, 2022.

[73] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. Dynamic video segmentation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6556–6565, 2018.

[74] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[75] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[76] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020.

[77] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context for semantic segmentation. *International Journal of Computer Vision*, 129(8):2375–2398, 2021.

[78] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34: 10326–10338, 2021.

[79] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2339–2348, 2022.

[80] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[81] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018.

[82] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[83] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[84] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017.

[85] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.