RankSeg: Adaptive Pixel Classification with Image Category Ranking for Segmentation

Haodi He[1†], Yuhui Yuan[3†], Xiangyu Yue[2], and Han Hu[3]

[1] University of Science and Technology of China
[2] UC Berkeley
[3] Microsoft Research Asia
✉ {yuhui.yuan,hanhu}@microsoft.com

Abstract. The segmentation task has traditionally been formulated as a complete-label[4] pixel classification task to predict a class for each pixel from a fixed number of predefined semantic categories shared by all images or videos. Yet, following this formulation, standard architectures will inevitably encounter various challenges under more realistic settings where the scope of categories scales up (e.g., beyond the level of 1k). On the other hand, in a typical image or video, only a few categories, i.e., a small subset of the complete label are present. Motivated by this intuition, in this paper, we propose to decompose segmentation into two sub-problems: (i) image-level or video-level multi-label classification and (ii) pixel-level rank-adaptive selected-label classification. Given an input image or video, our framework first conducts multi-label classification over the complete label, then sorts the complete label and selects a small subset according to their class confidence scores. We then use a rank-adaptive pixel classifier to perform the pixel-wise classification over only the selected labels, which uses a set of rank-oriented learnable temperature parameters to adjust the pixel classifications scores. Our approach is conceptually general and can be used to improve various existing segmentation frameworks by simply using a lightweight multi-label classification head and rank-adaptive pixel classifier. We demonstrate the effectiveness of our framework with competitive experimental results across four tasks, including image semantic segmentation, image panoptic segmentation, video instance segmentation, and video semantic segmentation. Especially, with our RankSeg, Mask2Former gains +0.8%/+0.7%/+0.7% on ADE20K panoptic segmentation/YouTubeVIS 2019 video instance segmentation/VSPW video semantic segmentation benchmarks respectively. Code is available at: `https://github.com/openseg-group/RankSeg`

Keywords: Rank-Adaptive, Selected-Label, Image Semantic Segmentation, Image Panoptic Segmentation, Video Instance Segmentation, Video Semantic Segmentation

---

[†] Equal contribution.

[4] We use the term "complete label" to represent the set of all predefined categories in the dataset.

(a) Image   (b) Ground-truth   (c) Multi-label image classification prediction
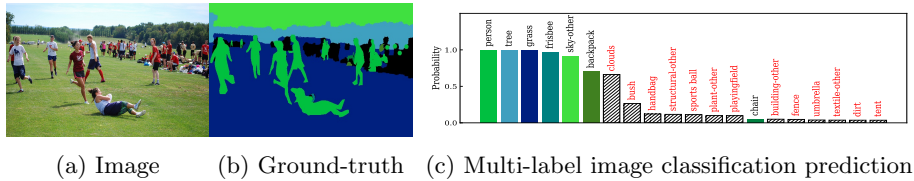
Fig. 1: **Illustrating the motivation of exploiting the multi-label image classification**: (a) An example image selected from COCO-Stuff. (b) The ground-truth segmentation map consisting of 7 classes. (c) The histogram of existence probability over the selected top 20 categories sorted by their confidence scores, which are predicted with our method. Our method only needs to identify the label of each pixel from these selected 20 categories instead of all 171 categories. The names of true/false positive categories are marked with black/red color, respectively. The bars associated with true positive categories share the same color as the ones adopted in the ground-truth segmentation map.

1   Introduction

Image and video segmentation, i.e., partitioning images or video frames into multiple meaningful segments, is a fundamental computer vision research topic that has wide applications including autonomous driving, surveillance system, and augmented reality. Most recent efforts have followed the path of fully convolutional networks [54] and proposed various advanced improvements, e.g., high-resolution representation learning [60,70], contextual representation aggregation [12,23,35,87,81], boundary refinement [66,43,84], and vision transformer architecture designs [53,61,82,88].

Most of the existing studies formulate the image and video segmentation problem as a complete-label pixel classification task. In the following discussion, we'll take image semantic segmentation as an example for convenience. For example, image semantic segmentation needs to select the label of each pixel from the complete label[5] set that is predefined in advance. However, it is unnecessary to consider the complete label set for every pixel in each image as most standard images only consist of objects belonging to a few categories. Figure 2 plots the statistics on the percentage of images that contain no more than the given class number in the entire dataset vs. the number of classes that appear within each image. Accordingly, we can see that 100.00%, 99.99%, 99.14%, and 99.85% of images contain less than 25 categories on PASCAL-Context [57], COCO-Stuff [8], ADE20K-Full [19,89], and COCO+LVIS [29,39] while each of them contains 60, 171, 847, and 1,284 predefined semantic categories respectively. Besides, Figure 1 shows an example image that only contains 7 classes while the complete label set consists of 171 predefined categories.

To take advantage of the above observations, we propose to re-formulate the segmentation task into two sub-problems including multi-label image/video classification and rank-adaptive pixel classification over a subset of selected labels.

---

[5] We use "label", "category", and "class" interchangeably.

(a) PASCAL-Context [57]   (b) COCO-Stuff [8]   (c) ADE20K-Full [89,19]   (d) COCO+LVIS [29,39]
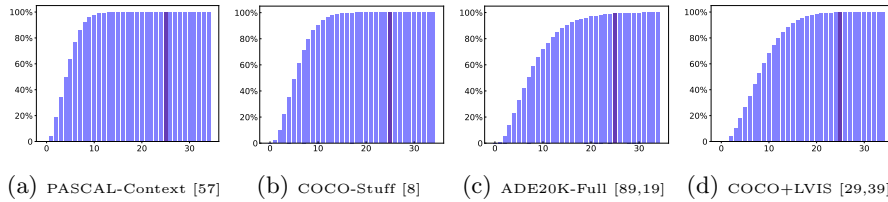
Fig. 2: **Illustrating the cumulative distribution of the number of images that contain no more than the given class number**: The x-axis represents the class number presented in the image and the y-axis represents the percentage of images that contain no more than the given class number in the entire dataset. We plot the cumulative distribution on four benchmarks. We can see that more than 99% of all images only contain less than 25 categories on four benchmarks, which are marked with dark purple bars. The above four benchmarks contain 60, 171, 847, and 1, 284 predefined semantic categories respectively.

(a) Image semantic seg.   (b) Image panoptic seg.   (c) Video semantic seg.
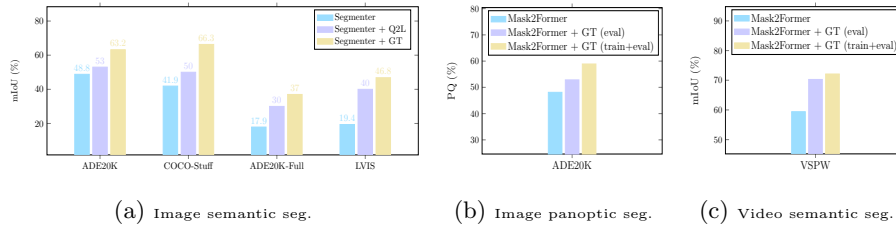
Fig. 3: **Illustrating the effectiveness of multi-label classification:**   The segmentation results of exploiting the ground-truth multi-label during only evaluation or both training and evaluation on (a) PASCAL-Context, COCO-Stuff, ADE20K-Full, and COCO+LVIS (image semantic segmentation)/(b) ADE20K (image panoptic segmentation)/(c) VSPW (video semantic segmentation). Refer to Sec.3.4 for more details. YouTubeVIS results are not included as we can not access the ground-truth.

To verify the potential benefits of our method, we investigate the gains via exploiting the ground-truth multi-label of each image or video. In other words, the pixel classifier only needs to select the category of each pixel from a collection of categories presented in the current image or video, therefore, we can filter out all other categories that do not appear. Figure 3 summarizes the comparison results based on Segmenter [65] and Mask2Former [18]. We can see that the segmentation performance is significantly improved given the ground-truth multi-label prediction. In summary, multi-label classification is an important but long-neglected sub-problem on the path toward more accurate segmentation.

Motivated by the significant gains obtained with the ground-truth multi-label, we propose two different schemes to exploit the benefit of multi-label image predictions including the *independent single-task scheme* and *joint multi-task scheme*. For the independent single-task scheme, we train one model for multi-label image/video classification and another model for segmentation. Specifically, we first train a model to predict multi-label classification probabilities for each

image/video, then we estimate the existing label subset for each image/video based on the multi-label predictions, last we use the predicted label subset to train the segmentation model for rank-adaptive selected-label pixel classification during both training and testing. For the joint multi-task scheme, we train one model to support both multi-label image/video classification and segmentation based on a shared backbone. Specifically, we apply a multi-label prediction head and a segmentation head, equipped with a rank-adaptive adjustment scheme, over the shared backbone and train them jointly. For both schemes, we need to send the multi-label predictions into the segmentation head for rank-adaptive selected-label pixel classification, which enables selecting a collection of categories that appear and adjusting the pixel-level classification scores according to the image/video content adaptively.

We demonstrate the effectiveness of our approach on various strong baseline methods including DeepLabv3 [12], Segmenter [65], Swin-Transformer [53], BEiT [4], MaskFormer [19], Mask2Former [18], and ViT-Adapter [15] across multiple segmentation benchmarks including PASCAL-Context [57], ADE20K [89], COCO-Stuff [8], ADE20K-Full [89,19], COCO+LVIS [29,39], YouTubeVIS [78], and VSPW [56].

2   Related Work

Image segmentation. We can roughly categorize the existing studies on image semantic segmentation into two main paths: (i) region-wise classification methods [1,7,27,26,74,58,7,67], which first organize the pixels into a set of regions (usually super-pixels), and then classify each region to get the image segmentation result. Several very recent methods [19,86,69] exploit the DETR framework [10] to conduct region-wise classification more effectively; (ii) pixel-wise classification methods, which predict the label of each pixel directly and dominate most previous studies since the pioneering FCN [54]. There exist extensive follow-up studies that improve the pixel classification performance via constructing better contextual representations [87,12,83,81] or designing more effective decoder architectures [3,64,13]. Image panoptic segmentation [42,41] aims to unify image semantic segmentation and image instance segmentation tasks. Some recent efforts have introduced various advanced architectures such as Panoptic FPN [41], Panoptic DeepLab [17], Panoptic Segformer [48], K-Net [86], and Mask2Former [18]. Our RankSeg is complementary with various paradigms and consistently improves several representative state-of-the-art methods across both image semantic segmentation and image panoptic segmentation tasks.

Video segmentation. Most of the previous works address the video segmentation task by extending the existing image segmentation models with temporal consistency constraint [71]. Video semantic segmentation aims to predict the semantic category of all pixels in each frame of a video sequence, where the main efforts focus on two paths including exploiting cross-frame relations to improve the prediction accuracy [44,36,40,25,59,11] and leveraging the information of neighboring frames to accelerate computation [55,77,47,38,34,52]. Video instance seg-

mentation [78] requires simultaneous detection, segmentation and tracking of instances in videos and there exist four mainstream frameworks including tracking-by-detection [68,49,9,24,32], clip-and-match [6,2], propose-and-reduce [50], and segment-as-a-whole [72,37,75,16]. We show the effectiveness of our method on both video semantic segmentation and video instance segmentation tasks via improving the very recent state-of-the-art method Mask2Former [16].

Multi-label classification. The goal of multi-label classification is to identity all the categories presented in a given image or video over the complete label set. The conventional multi-label image classification literature partitions the existing methods into three main directions: (i) improving the multi-label classification loss functions to handle the imbalance issue [76,5], (ii) exploiting the label co-occurrence (or correlations) to model the semantic relationships between different categories [33,46,14,79], and (iii) localizing the diverse image regions associated with different categories [73,28,80,45,51]. In our independent single-task scheme, we choose the very recent state-of-the-art method Query2Label [51] to perform multi-label classification on various semantic segmentation benchmarks as it is a very simple and effective method that exploits the benefits of both label co-occurrence and localizing category-dependent regions. There exist few efforts that apply multi-label image classification to address segmentation task. To the best of our knowledge, the most related study EncNet [85] simply adds a multi-label image classification loss w/o changing the original semantic segmentation head that still needs to select the label of each pixel from all predefined categories. We empirically show the advantage of our method over EncNet in the ablation experiments. Besides, our proposed method is naturally suitable to solve large-scale semantic segmentation problem as we only perform rank-adaptive pixel classification over a small subset of the complete label set based on the multi-label image prediction. We also empirically verify the advantage of our method over the very recent ESSNet [39] in the ablation experiments.

3 Our Approach

We first introduce the overall framework of our approach in Sec. 3.1, which is also illustrated in Figure 4. Second, we introduce the details of the independent single-task scheme in Sec. 3.2 and those of joint multi-task scheme in Sec. 3.3. Last, we conduct analysis experiments to investigate the detailed improvements of our method across multiple segmentation tasks in Sec. 3.4.

3.1 Framework

The overall framework of our method is illustrated in Figure 4, which consists of one path for multi-label image classification and one path for semantic segmentation. The multi-label prediction is used to sort and select the top κ category embeddings with the highest confidence scores which are then sent into the rank-adaptive selected-label pixel classifier to generate the semantic segmentation prediction. We explain the mathematical formulations of both multi-label
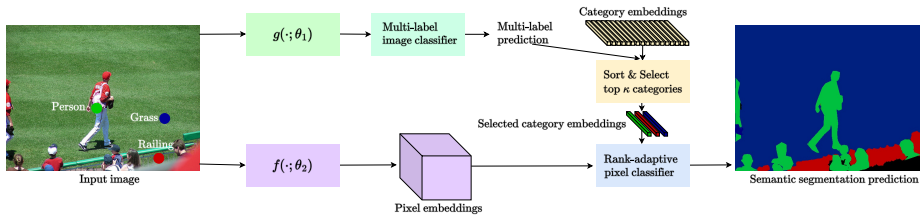
Fig. 4: **Illustrating the framework of our approach:** Given an input image that contains person, grass, and railing, which are marked with ●, ●, and ●, respectively. First, we use the multi-label image classification model $g(\cdot; \theta_1)$ and multi-label image classifier to predict the presence probabilities of all categories. Second, we sort the labels according to the predicted presence probabilities and select the top κ most probable category embeddings. Last, we send the selected subset of category embeddings into the rank-adaptive selected-label pixel classifier to identify the label of each pixel based on the pixel embeddings output by the semantic segmentation model $f(\cdot; \theta_2)$.

image classification and rank-adaptive selected-label pixel classification as follows:

Multi-label image classification. The goal of multi-label image classification is to predict the set of existing labels in a given image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the input height and width. We generate the image-level multi-label ground truth $\mathbf{y}^{gt}$ from the ground truth segmentation map and we represent $\mathbf{y}^{gt}$ with a vector of K binary values $[y_1^{gt}, y_2^{gt}, ..., y_K^{gt}]^\mathsf{T}, y_i^{gt} \in \{0, 1\}$, where K represents the total number of predefined categories and $y_i^{gt} = 1$ represents the existence of pixels belonging to i-th category and $y_i^{gt} = 0$ otherwise.

The prediction $\mathbf{y} \in \mathbb{R}^K$ is a vector that records the existence confidence score of each category in the given image $\mathbf{x}$. We use $g(\cdot; \theta_1)$ to represent the backbone for the multi-label image classification model. We estimate the multi-label predictions of input $\mathbf{x}$ with the following sigmoid function:

$$y_k = \frac{e^{\xi(g(\mathbf{x};\theta_1), \mathbf{h}_k)}}{e^{\xi(g(\mathbf{x};\theta_1), \mathbf{h}_k)} + 1}, \tag{1}$$

where y_k is the k-th element of $\mathbf{y}$, $g(\mathbf{x}; \theta_1)$ represents the output feature map, $\mathbf{h}_k$ represents the multi-label image classification weight associated with the k-th category, and $\xi(\cdot)$ represents a transformation function that estimates the similarity between the output feature map and the multi-label image classification weights. We supervise the multi-label predictions with the asymmetric loss that operates differently on positive and negative samples by following [5,51].

Rank-adaptive selected-label pixel classification. The goal of semantic segmentation is to predict the semantic label of each pixel and the label is selected from all predefined categories. We use $\mathbf{z} \in \mathbb{R}^{H \times W \times K}$ to represent the predicted pixel classification probability map for the input image $\mathbf{x}$. We use $f(\cdot; \theta_2)$ to represent the semantic segmentation backbone and $\mathbf{z}^{gt} \in \mathbb{R}^{H \times W}$ to represent the ground-truth segmentation map. Instead of choosing the label of
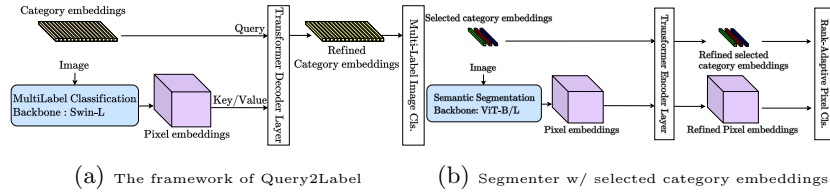
(a) The framework of Query2Label   (b) Segmenter w/ selected category embeddings

Fig. 5: (a) **Illustrating the framework of Query2Label**: The multi-label image classification backbone $g(\cdot; \theta_1)$ is set as Swin-L by default. The transformation $\xi(\cdot)$ is implemented as two transformer decoder layers followed by a linear layer that prepares the refined category embeddings for the multi-label image classifier. (b) **Illustrating the framework of Segmenter w/ selected category embeddings:** The semantic segmentation backbone $f(\cdot; \theta_2)$ is set as ViT-B/16 or ViT-L/16. The transformation $\psi(\cdot)$ is implemented as two transformer encoder layers followed by ℓ_2-normalization before estimating the segmentation map.

each pixel from all K predefined categories, based on the previous multi-label prediction $\mathbf{y}$ for image $\mathbf{x}$, we introduce a more effective rank-adaptive selected-label pixel classification scheme:

- Sort and select the top κ elements of the classifier weights according to the descending order of multi-label predictions $\mathbf{y} = [y_1, y_2, \cdots, y_K]$:

$$[\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2, \cdots, \overline{\mathbf{w}}_\kappa] = \text{Top-}\kappa([\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K], \mathbf{y}), \qquad (2)$$

- Rank-adaptive classification of pixel (i, j) over the top κ selected categories:

$$\mathbf{z}_{i,j,k} = \frac{e^{\psi(f(\mathbf{x};\theta_2)_{i,j}, \overline{\mathbf{w}}_k)/\tau_k}}{\sum_{l=1}^{\kappa} e^{\psi(f(\mathbf{x};\theta_2)_{i,j}, \overline{\mathbf{w}}_l)/\tau_l}}, \qquad (3)$$

where $[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K]$ represents the pixel classification weights for all K predefined categories and $[\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2, \cdots, \overline{\mathbf{w}}_\kappa]$ represents the top κ selected pixel classification weights associated with the largest multi-label classification scores. $f(\mathbf{x};\theta_2)$ represents the output feature map for semantic segmentation. $\psi(\cdot)$ represents a transformation function that estimates the similarity between the pixel features and the pixel classification weights. κ represents the number of selected category embeddings and κ is chosen as a much smaller value than K. We apply a set of rank-adaptive learnable temperature parameters $[\tau_1, \tau_2, \cdots, \tau_\kappa]$ to adjust the classification scores over the selected top κ categories. The temperature parameters across different selected classes are shared in all of the baseline experiments by default[6]. We analyze the influence of κ choices and the benefits of such a rank-oriented adjustment scheme in the following discussions and experiments.

---

[6] We set $\tau_1 = \tau_2 = \cdots = \tau_\kappa$ for all baseline segmentation experiments.
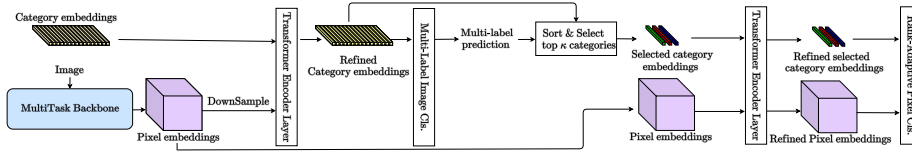
Fig. 6: **Illustrating the framework of joint multi-task scheme:** $g(\cdot; \theta_1)$ and $f(\cdot; \theta_2)$ are set as the shared multi-task backbone and $\theta_1 = \theta_2$. $\xi(\cdot)$ is implemented as one transformer encoder layer or two transformer decoder layers or global average pooling + linear projection. $\psi(\cdot)$ is implemented as two transformer encoder layers followed by L2-normalization before estimating the segmentation map.

3.2   Independent single-task scheme

Under the independent single-task setting, the multi-label image classification model $g(\cdot; \theta_1)$ and the semantic segmentation model $f(\cdot; \theta_2)$ are trained separately and their model parameters are not shared, i.e., $\theta_1 \neq \theta_2$. Specifically, we first train the multi-label image classification model $g(\cdot; \theta_1)$ to identify the top κ most likely categories for each image. Then we train the rank-adaptive selected-label pixel classification model, i.e., semantic segmentation model, $f(\cdot; \theta_2)$ to predict the label of each pixel over the selected top κ classes.

Multi-label image classification model. We choose the very recent SOTA multi-label classification method Query2Label [51] as it performs best on multiple multi-label classification benchmarks by the time of our submission according to paper-with-code.[7] The key idea of Query2Label is to use the category embeddings as the query to gather the desired pixel embeddings as the key/value, which is output by an ImageNet-22K pre-trained backbone such as Swin-L, adaptively with one or two transformer decoder layers. Then Query2Label scheme applies a multi-label image classifier over the refined category embeddings to predict the existence of each category. Figure 5a illustrates the framework of Query2Label framework. Refer to [51] and the official implementation for more details. The trained weights of Query2Label model are fixed during both training and inference of the following semantic segmentation model.

Rank-adaptive selected-label pixel classification model. We choose a simple yet effective baseline Segmenter [65] as it achieves even better performance than Swin-L[8] when equipped with ViT-L. Segmenter first concatenates the category embedding with the pixel embeddings output by a ViT model together and then sends them into two transformer encoder layers. Last, based on the refined pixel embeddings and category embeddings, Segmenter computes their ℓ_2-normalized scalar product as the segmentation predictions. We select the top κ most likely categories for each image according to the predictions of the Query2Label model and only use the selected top κ category embeddings

---

[7] https://paperswithcode.com/task/multi-label-classification
[8] Segmenter w/ ViT-L: 53.63% vs. Swin-L: 53.5% on ADE20K.

instead of all category embeddings. Figure 5b illustrates the overall framework of Segmenter with the selected category embeddings.

3.3   Joint multi-task scheme

Considering that the independent single-task scheme suffers from extra heavy computation overhead as the Query2Label method relies on a large backbone, e.g., Swin-L, we introduce a joint multi-task scheme that shares the backbone for both sub-tasks, in other words, $\theta_1 = \theta_2$ and the computations of $g(\cdot; \theta_1)$ and $f(\cdot; \theta_2)$ are also shared.

Figure 6 shows the overall framework of the joint multi-task scheme. First, we apply a shared multi-task backbone to process the input image and output the pixel embeddings. Second, we concatenate the category embeddings with the down-sampled pixel embeddings[9], send them into one transformer encoder layer, and apply the multi-label image classifier on the refined category embeddings to estimate the multi-label predictions. Last, we sort and select the top κ category embeddings, concatenate the selected category embeddings with the pixel embeddings, send them into two transformer encoder layers, and compute the semantic segmentation predictions based on ℓ_2-normalized scalar product between the refined selected category embeddings and the refined pixel embeddings. We empirically verify the advantage of the joint multi-task scheme over the independent single-task scheme in the ablation experiments.

3.4   Analysis experiments

Oracle experiments. We first conduct several groups of oracle experiments based on Segmenter w/ ViT-B/16 on four challenging image semantic segmentation benchmarks (PASCAL-Context/COCO-Stuff/ADE20K-Full/COCO+LVIS), Mask2Former w/ Swin-L on both ADE20K panoptic segmentation benchmark[10] and VSPW video semantic segmentation benchmark.
- *Segmenter/Mask2Former + GT (train + eval)*: the upper-bound segmentation performance of Segmenter/Mask2Former when training & evaluating equipped with the ground-truth multi-label of each image or video, in other words, we only need to select the category of each pixel over the ground-truth existing categories in a given image or video.
- *Segmenter/Mask2Former + GT (eval)*: the upper-bound segmentation performance of Segmenter/Mask2Former when only using the ground-truth multi-label of each image or video during evaluation.

Figure 3 illustrates the detailed comparison results. We can see that only applying the ground-truth multi-label during evaluation already brings considerable improvements and further applying the ground-truth multi-label during

---

[9] Different from the semantic segmentation task, the multi-label image classification task does not require high-resolution representations.

[10] We choose Swin-L by following the MODEL_ZOO of the official Mask2Former implementation: https://github.com/facebookresearch/Mask2Former

Table 1: Ablation of the improvements with our method. MT: multi-task learning with auxiliary multi-label image classification scheme. LS: label sort and selection, in other words, sort and select the top κ classes. RA: rank-adaptive-τ, which applies independent τ for pixel classification scores associated with the different ranking positions.

Method.	Image semantic seg.				Image panoptic seg.	Video semantic seg.
	PASCAL-Context	COCO-Stuff	ADE20K-Full	COCO+LVIS	ADE20K	VSPW
Baseline	53.85	41.85	17.93	19.41	48.1	59.4
+ MT	54.05	42.38	17.81	20.26	48.2	59.5
+ MT + LS	54.27	44.31	18.26	21.13	48.8	59.6
+ MT + LS + RA	**54.76**	**44.98**	**18.78**	**21.26**	**48.9**	**60.1**

training significantly improves the segmentation performance across all benchmarks. For example, when compared to the baseline Segmenter or Mask2Former, Segmenter + GT (train + eval) gains +17%/+24%/+19%/+27% absolute mIoU scores across PASCAL-Context/COCO-Stuff/ADE20K-Full/COCO+LVIS and Mask2Former + GT (train + eval) gains +11%/+13% on ADE20K/VSPW.

Improvement analysis of RankSeg. Table 1 reports the results with different combinations of the proposed components within our joint multi-task scheme. We can see that: (i) multi-task learning (MT) introduces the auxiliary multi-label image classification task and brings relatively minor gains on most benchmarks, (ii) combining MT with label sorting & selection (LS) achieves considerable gains, and (iii) applying the rank-adaptive-τ manner (shown in Equation 3) instead of shared-τ achieves better performance. We investigate the possible reasons by analyzing the value distribution of learned $1/\tau$ with the rank-adaptive-τ manner in Figure 7, which shows that the learned $1/\tau$ is capable of adjusting the pixel classification scores based on the order of multi-label classification scores. In summary, we choose "MT + LS + RA" scheme by default, which gains +0.91%/+3.13%/+0.85%/+1.85%/+0.8%/+0.7% over the baseline methods across these six challenging segmentation benchmarks respectively.

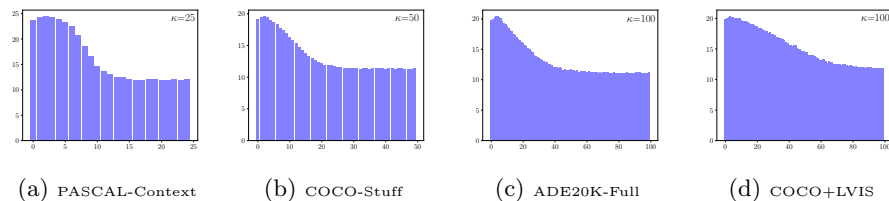(a) PASCAL-Context   (b) COCO-Stuff   (c) ADE20K-Full   (d) COCO+LVIS

Fig. 7: Illustrating the values of learned $1/\tau$ with rank-adaptive-τ manner on four semantic segmentation benchmarks. We can see that the values of $1/\tau$ are almost monotonically decreasing, thus meaning that the pixel classification scores of the category associated with larger multi-label classification scores are explicitly increased.

4   Experiment

4.1   Implementation details

We illustrate the details of the datasets, including ADE20K [89], ADE20K-Full [89,19], PASCAL-Context [57], COCO-Stuff [8], COCO+LVIS [29,39], VSPW [56], and YouTubeVIS [78], in the supplementary material.

Multi-label image classification. For the independent single-task scheme, following the official implementation[11] of Query2Label, we train multi-label image classification models, e.g., ResNet-101 [31], TResNetL [62], and Swin-L [53], for 80 epochs using Adam solver with early stopping. Various advanced tricks such as cutout [22], RandAug [21] and EMA [30] are also used. For the joint multi-task scheme, we simply train the multi-label image classification models following the same settings as the segmentation models w/o using the above-advanced tricks that might influence the segmentation performance. We illustrate more details of the joint multi-task scheme in the supplementary material.

Segmentation. We adopt the same settings for both independent single-task scheme and joint multi-task scheme. For the segmentation experiments based on Segmenter [65], DeepLabv3 [12], Swin-Transformer [53], BEiT [4], and ViT-Adapter [15]. we follow the default training & testing settings of their reproduced version based on `mmsegmentation` [20]. For the segmentation experiments based on MaskFormer or Mask2Former, we follow their official implementation[12].

Hyper-parameters. We set κ as 25, 50, 50, 100, and 100 on PASCAL-Context, ADE20K, COCO-Stuff, ADE20K-Full, and COCO+LVIS respectively as they consist of a different number of semantic categories. We set their multi-label image classification loss weights as 5, 10, 10, 100, and 300. The segmentation loss weight is set as 1. We illustrate the hyper-parameter settings of experiments on MaskFormer [19] or Mask2Former [18] in the supplementary material.

Metrics. We report mean average precision (mAP) for multi-label image classification task, mean intersection over union (mIoU) for image/video semantic segmentation task, panoptic quality (PQ) for panoptic segmentation task, and mask average precision (AP) for instance segmentation task.

4.2   Ablation experiments

We conduct all ablation experiments based on the Segmenter w/ ViT-B and report their single-scale evaluation results on COCO-Stuff `test` and COCO+LVIS `test` if not specified. The baseline with Segmenter w/ ViT-B achieves 41.85% and 19.41% mIoU on COCO-Stuff and COCO+LVIS, respectively.

Influence of the multi-label classification accuracy. We investigate the influence of multi-label classification accuracy on semantic segmentation tasks based on the independent single-task scheme. We train multiple Query2Label models based on different backbones, e.g., ResNet101, TResNetL, and Swin-L.

---

[11] https://github.com/SlongLiu/query2labels
[12] https://github.com/facebookresearch/Mask2Former

Table 2: Influence of the multi-label image classification accuracy (mAP) on semantic segmentation accuracy (mIoU) based on independent single-task manner.

Backbone	COCO-Stuff			COCO+LVIS		
	ResNet101 [63]	TResNetL [63]	Swin-L [53]	ResNet101 [63]	TResNetL [63]	Swin-L [53]
mAP (%)	55.17	60.10	**64.79**	26.54	31.01	**34.93**
mIoU (%)	39.34	42.87	**44.42**	16.30	19.88	**21.19**
△	-2.51	+1.02	**+2.57**	-3.11	+0.47	**+1.78**

Table 3: Independent single-task scheme vs. Joint multi-task scheme: we adopt Swin-L as the backbone for the multi-label predictions in the independent single-task scheme.

Method	COCO-Stuff				COCO+LVIS			
	#params.	FLOPs	mIoU (%)	△	#params.	FLOPs	mIoU (%)	△
Indep. single-task.	343.21M	182.8G	44.42	+2.52	347.33M	233.62G	21.19	+1.78
Joint multi-task.	**109.73**M	**78.71**G	**44.98**	**+3.08**	**111.45**M	**99.59**G	**21.26**	**+1.85**

Table 4: Influence of the size of the selected label set, i.e., κ.

κ	COCO-Stuff							COCO+LVIS			
	25	50	75	100	125	150	171	50	100	200	1,284
mIoU (%)	44.65	**44.98**	44.58	44.67	44.41	44.37	44.20	20.36	**21.26**	20.99	20.93
△	+2.80	**+3.13**	+2.73	+2.82	+2.56	+2.52	+2.35	+0.95	**+1.85**	+1.58	+1.52

Table 5: Influence of the multi-label classification loss weight.

multi-label cls. loss weight	1	5	10	20
mAP (%)	59.14	62.38	**62.52**	62.18
mIoU (%)	44.11	44.90	**44.98**	44.04
△	+2.26	+3.05	**+3.13**	+2.19

Table 6: Influence of multi-label prediction head architecture.

Method	#params.	FLOPs	mAP	mIoU (%)	△
GAP+Linear	**103.24**M	**76.94**G	60.62	43.19	+1.34
2× TranDec	114.27M	78.83G	61.15	44.07	+2.22
1× TranEnc	109.73M	78.71G	**62.52**	**44.98**	**+3.13**

Then we train three Segmenter w/ ViT-B segmentation models based on their multi-label predictions independently. According to the results in Table 2, we can see that more accurate multi-label classification prediction brings more semantic segmentation performance gains and less accurate multi-label classification prediction even results in worse results than baseline.

Independent single-task scheme vs. Joint multi-task scheme. We compare the performance and model complexity of the independent single-task scheme (w/ Swin-L) and joint multi-task scheme in Table 3. To ensure fairness, we choose the # of labels in the selected label set, i.e., κ, as 50 for both schemes, and the segmentation models are trained & tested under the same settings. According to the comparison results, we can see that joint multi-task scheme achieves better performance with fewer parameters and FLOPs. Thus, we choose the joint multi-task scheme in the following experiments for efficiency if not specified.

Influence of different top κ. We study the influence of the size of selected label set, i.e., κ, as shown in Table 4. According to the results, our method achieves the best performance on COCO-Stuff/COCO+LVIS when κ=50/κ=100, which achieves a better trade-off between precision and recall for multi-label predictions. Besides, we attempt to fix κ=50 during training and report the results when changing κ during evaluation on COCO-Stuff: κ=25: 44.89% /κ=50:

44.98% /κ=75: 45.00% /κ=100: 45.01%. Notably, we also report the results with κ=K, in other words, we only sort the classifier weights, which also achieve considerable gains. Therefore, we can see that sorting the classes and rank-adaptive adjustment according to multi-label predictions are the key to the gains. In summary, our method consistently outperforms baseline with different κ values. We also attempt to use dynamic κ for different images during evaluation but observe no significant gains. More details are provided in the supplementary material.

Influence of the multi-label classification loss weight. We study the influence of the multi-label image classification loss weights with the joint multi-task scheme on COCO-Stuff and report the results in Table 5. We can see that setting the multi-label classification loss weight as 10 achieves the best performance.

Influence of $\xi(\cdot)$ choice. Table 6 compares the results based on different multi-label prediction head architecture choices including "GAP+Linear" (applying global average pooling followed by linear projection), "2$\times$ TranDec" (using two transformer decoder layers), and "1$\times$ TranEnc" (using one transformer encoder layer) on COCO-Stuff. Both "2$\times$ TranDec" and "1$\times$ TranEnc" operate on feature maps with $\frac{1}{32}$ resolution of the input image. According to the results, we can see that "1$\times$ TranEnc" achieves the best performance and we implement $\xi(\cdot)$ as one transformer encoder layer if not specified. We also attempt generating multi-label prediction (mAP=58.55%) from the semantic segmentation prediction directly but observe no performance gains, thus verifying the importance of relatively more accurate multi-label classification predictions.

More comparison results with the previous EncNet [85] and ESSNet [39] are summarized in the supplementary material.

4.3 State-of-the-art experiments

Image semantic segmentation. We apply our method to various state-of-the-art image semantic segmentation methods including DeepLabv3, Seg-Mask-L/16, Swin-Transformer, BEiT, and ViT-Adapter-L. Table 7 summarizes the detailed comparison results across three semantic segmentation benchmarks including ADE20K, COCO-Stuff, and COCO+LVIS, where we evaluate the multi-scale segmentation results on ADE20K/COCO-Stuff and single-scale segmentation results on COCO+LVIS (due to limited GPU memory) respectively. More details of how to apply our joint multi-task scheme to these methods are provided in the supplementary material. According to the results in Table 7, we can see that our RankSeg consistently improves DeepLabv3, Seg-Mask-L/16, Swin-Transformer, BEiT, and ViT-Adapter-L across three evaluated benchmarks. For example, with our RankSeg, BEiT gains 0.8% on ADE20K with slightly more parameters and GFLOPs.

Image panoptic segmentation & Video semantic segmentation & Video instance segmentation. To verify the generalization ability of our method, we extend RankSeg to "rank-adaptive selected-label region classification" and apply it to the very recent Mask2Former [18]. According to Table 8, our RankSeg improves the image semantic segmentation/image panoptic segmentation/video semantic segmentation/video instance segmentation performance by +0.7%/+0.8%

Table 7: Combination with DeepLabv3, Seg-Mask-L, Swin-B, BEiT, and ViT-Adapter.

Method	ADE20K			COCO-Stuff			COCO+LVIS		
	#params.	FLOPs	mIoU(%)	#params.	FLOPs	mIoU(%)	#params.	FLOPs	mIoU(%)
DeepLabv3	**87.21**M	**347.64**G	45.19	**87.22**M	**347.68**G	38.42	**88.08**M	**350.02**G	11.04
+ RankSeg	91.87M	349.17G	**46.61**	91.75M	349.09G	**39.86**	93.21M	359.47G	**12.76**
Seg-Mask-L/16	**333.23**M	377.83G	53.63	**333.26**M	378.57G	47.12	**334.4**M	**420.2**G	23.71
+ RankSeg	345.99M	**377.18**G	**54.47**	346.03M	**377.46**G	**47.93**	348.31M	422.6G	**24.60**
Swin-B	**121.42**M	**299.81**G	52.4	**121.34**M	**299.98**G	47.16	**122.29**M	309.34G	20.33
+ RankSeg	125.43M	300.48G	**53.01**	125.46M	300.56G	**47.85**	126.89M	**306.53**G	**20.81**
BEiT	**441.27**M	**1745.99**G	57.0	**441.30**M	**1746.54**G	49.9		OOM	
+ RankSeg	456.28M	1751.04G	**57.8**	456.35M	1751.34G	**50.3**		OOM	
ViT-Adapter-L	**570.74**M	**2743.20**G	60.5		N/A			N/A	
+ RankSeg	584.71M	2747.81G	**60.7**		N/A			N/A	

OOM means out of memory error on 8×32G V100 GPUs.

N/A means the results are available due to limited GPU resources.

Table 8: Combination with Mask2Former based on Swin-L.

Method	Image semantic seg.			Image panoptic seg.			Video semantic seg.			Video instance seg.		
	ADE20K			ADE20K			VSPW (T=2)			YouTubeVIS 2019 (T=2)		
	#params.	FLOPs	mIoU(%)	#params.	FLOPs	PQ(%)	#params.	FLOPs	mIoU(%)	#params.	FLOPs	AP(%)
Mask2Former [18]	**205.51**M	**369.02**G	57.3	**205.55**M	**377.98**G	48.1	**205.50**M	**737.38**G	59.4	**205.52**M	**753.79**G	60.4
+ RankSeg	208.79M	369.83G	**58.0**	208.83M	379.11G	**48.9**	208.75M	738.14G	**60.1**	208.70M	754.59G	**61.1**

T=2 means each video clip is composed of 2 frames during training/evaluation and we report the GFLOPs over 2 frames.

/+0.7%/+0.7% respectively with slightly more parameters and GFLOPs. More details about the Mask2Former experiments and the results of combining Mask-Former with RankSeg are provided in the supplementary material.

5 Conclusion

This paper introduces a general and effective rank-oriented scheme that formulates the segmentation task into two sub-problems including multi-label classification and rank-adaptive selected-label pixel classification. We first verify the potential benefits of exploiting multi-label image/video classification to improve pixel classification. We then propose a simple joint multi-task scheme that is capable of improving various state-of-the-art segmentation methods across multiple benchmarks. We hope our initial attempt can inspire more efforts towards using a rank-oriented manner to solve the challenging segmentation problem with a large number of categories. Last, we want to point out that designing & exploiting more accurate multi-label image/video classification methods is a long-neglected but very important sub-problem towards more general and accurate segmentation.

References

1. Arbeláez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., Malik, J.: Semantic segmentation using regions and parts. In: CVPR (2012)

2. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV. pp. 158–177. Springer (2020)

3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. PAMI (2017)

4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

5. Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. arXiv preprint arXiv:2009.14119 (2020)

6. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR. pp. 9739–9748 (2020)

7. Caesar, H., Uijlings, J., Ferrari, V.: Region-based semantic segmentation with end-to-end training. In: ECCV (2016)

8. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)

9. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV. pp. 1–18. Springer (2020)

10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)

11. Chandra, S., Couprie, C., Kokkinos, I.: Deep spatio-temporal random fields for efficient video segmentation. In: CVPR. pp. 8915–8924 (2018)

12. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)

13. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)

14. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: ICCV (2019)

15. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)

16. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)

17. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab. arXiv:1910.04751 (2019)

18. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arXiv preprint arXiv:2112.01527 (2021)

19. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. arXiv preprint arXiv:2107.06278 (2021)

20. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020)

21. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPRW. pp. 702–703 (2020)

22. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)

23. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154 (2019)

24. Fu, Y., Yang, L., Liu, D., Huang, T.S., Shi, H.: Compfeat: Comprehensive feature aggregation for video instance segmentation. arXiv preprint arXiv:2012.03400 **6** (2020)
25. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video cnns through representation warping. In: ICCV. pp. 4453–4462 (2017)
26. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
27. Gu, C., Lim, J.J., Arbelaez, P., Malik, J.: Recognition using regions. In: CVPR (2009)
28. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: CVPR (2019)
29. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019)
30. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
32. Hu, A., Kendall, A., Cipolla, R.: Learning a spatio-temporal embedding for video instance segmentation. arXiv preprint arXiv:1912.08969 (2019)
33. Hu, H., Zhou, G.T., Deng, Z., Liao, Z., Mori, G.: Learning structured inference neural networks with label relations. In: CVPR (2016)
34. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR. pp. 8818–8827 (2020)
35. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: CVPR. pp. 603–612 (2019)
36. Hur, J., Roth, S.: Joint optical flow and temporally consistent semantic segmentation. In: ECCV. pp. 163–177. Springer (2016)
37. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. NIPS **34** (2021)
38. Jain, S., Wang, X., Gonzalez, J.E.: Accel: A corrective fusion network for efficient semantic segmentation on video. In: CVPR. pp. 8866–8875 (2019)
39. Jain, S., Paudel, D.P., Danelljan, M., Van Gool, L.: Scaling semantic segmentation beyond 1k classes on a single gpu. In: ICCV. pp. 7426–7436 (2021)
40. Jin, X., Li, X., Xiao, H., Shen, X., Lin, Z., Yang, J., Chen, Y., Dong, J., Liu, L., Jie, Z., et al.: Video scene parsing with predictive feature learning. In: ICCV. pp. 5580–5588 (2017)
41. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: CVPR. pp. 6399–6408 (2019)
42. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR. pp. 9404–9413 (2019)
43. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR. pp. 9799–9808 (2020)
44. Kundu, A., Vineet, V., Koltun, V.: Feature space optimization for semantic video segmentation. In: CVPR. pp. 3168–3175 (2016)
45. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: CVPR (2021)
46. Li, Q., Qiao, M., Bian, W., Tao, D.: Conditional graphical lasso for multi-label image classification. In: CVPR (2016)
47. Li, Y., Shi, J., Lin, D.: Low-latency video semantic segmentation. In: CVPR. pp. 5997–6005 (2018)

48. Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Lu, T., Luo, P.: arXiv preprint arXiv:2109.03814 (2021)
49. Lin, C.C., Hung, Y., Feris, R., He, L.: Video instance segmentation tracking with a modified vae architecture. In: CVPR. pp. 13147–13157 (2020)
50. Lin, H., Wu, R., Liu, S., Lu, J., Jia, J.: Video instance segmentation with a propose-reduce paradigm. In: CVPR. pp. 1739–1748 (2021)
51. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)
52. Liu, Y., Shen, C., Yu, C., Wang, J.: Efficient semantic video segmentation with per-frame inference. In: ECCV. pp. 352–368. Springer (2020)
53. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
54. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
55. Mahasseni, B., Todorovic, S., Fern, A.: Budget-aware deep semantic video segmentation. In: CVPR. pp. 1029–1038 (2017)
56. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: 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. pp. 4133–4143 (2021)
57. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014)
58. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: CVPR (2017)
59. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: CVPR. pp. 6819–6828 (2018)
60. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: CVPR. pp. 4151–4160 (2017)
61. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
62. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses (2021)
63. Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., Friedman, I.: Tresnet: High performance gpu-dedicated architecture. In: WACV. pp. 1400–1409 (2021)
64. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
65. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. arXiv preprint arXiv:2105.05633 (2021)
66. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: ICCV. pp. 5229–5238 (2019)
67. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV (2013)
68. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: CVPR. pp. 7942–7951 (2019)
69. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: CVPR. pp. 5463–5474 (2021)
70. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. TPAMI (2019)

71. Wang, W., Zhou, T., Porikli, F., Crandall, D., Van Gool, L.: A survey on deep learning technique for video segmentation. arXiv preprint arXiv:2107.01153 (2021)
72. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR. pp. 8741–8750 (2021)
73. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: ICCV (2017)
74. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: CVPR (2017)
75. Wu, J., Jiang, Y., Zhang, W., Bai, X., Bai, S.: Seqformer: a frustratingly simple model for video instance segmentation. arXiv preprint arXiv:2112.08275 (2021)
76. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: ECCV (2020)
77. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: CVPR. pp. 6556–6565 (2018)
78. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV. pp. 5188–5197 (2019)
79. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: ECCV (2020)
80. You, R., Guo, Z., Cui, L., Long, X., Bao, Y., Wen, S.: Cross-modality attention with semantic graph embedding for multi-label classification. In: AAAI (2020)
81. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV. pp. 173–190. Springer (2020)
82. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution transformer for dense prediction. arXiv preprint arXiv:2110.09408 (2021)
83. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)
84. Yuan, Y., Xie, J., Chen, X., Wang, J.: Segfix: Model-agnostic boundary refinement for segmentation. In: ECCV (2020)
85. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: CVPR. pp. 7151–7160 (2018)
86. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. arXiv preprint arXiv:2106.14855 (2021)
87. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
88. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
89. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)