RWS: Refined Weak Slice for Semantic Segmentation Enhancement

Yunbo Rao, Qingsong Lv, Andrei Sharf, Zhanglin Cheng*

***Abstract*—Interpretation of predictions made by Convolutional Neural Networks (CNNs) is a rapidly growing field of research. A common approach involves enhancing semantic segmentation predictions through the generation of heatmaps that illustrate the significance of individual pixels in the segmentation. Nevertheless, the selection of beneficial features from these heatmaps remains a challenge. This is because the introduced information often contains interfering factors such as mutual features between different objects, background, and insufficient heat map resolution which often diminish its effectiveness. To overcome these limitations, we introduce Refined Weak Slices (RWS). Our main idea is to identify low attention regions in heat maps i.e. *weak slices*, in conjunction with segmentation accuracy, and utilize them to select effective features across different DNN layers, to enhance segmentation. We then seamlessly integrate these features back into the CNN, thus *refining* and enhancing the semantic segmentation result with selected features. Through extensive experiments, we demonstrate that incorporating the RWS module into state-of-the-art methods yields a notable improvement in the average mIoU by 2.84% on benchmark datasets (VOC 2012, COCOStuff, ADE20K, Cityscapes) for both ResNet-101 and ResNet-50 architectures. Furthermore, we achieve a maximum improvement of 5.8% with a single CNN. Overall, the combination of RWS and CNNs exhibits excellent performance in image segmentation tasks.**

***Index Terms*—Semantic Segmentation, Refine Slice Feature, Retraining.**

I. Introduction

In recent years, significant progress in semantic segmentation has been propelled by advancements in Convolutional Neural Networks (CNNs). CNN-based models have adopted diverse strategies, such as pyramid pooling [1], dilated convolutions [2], and self-attention [3]–[6], to improve segmentation accuracy. Nevertheless, understanding why and how CNNs achieve state-of-the-art results is still a challenging problem. Methods that aim to provide visual explanations for neural network decisions are particularly interested in unraveling the decision-making processes of CNNs, especially in the context of CNN enhancement.

This research was supported by the Science and Technology Project of Sichuan (No. 2021YFG0314, 2022ZHCG0033, 2023ZHCG0005, 2023ZHCG0008), and the National Natural Science Foundation of China (No. U19A2078, 61972388). (*Corresponding author: Zhanglin Cheng.*)

Y. Rao and Q. Lv are with the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (E-mail: raoyb@uestc.edu.cn and lvqingsong.0@outlook.com).

Andrei Sharf is with the Computer Science Department, Ben-Gurion University (E-mail: asharf@gamil.com).

Z. Cheng is with Shenzhen VisuCA Key Lab, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (zl.cheng@siat.ac.cn).

RWS code available at https://github.com/Lqs-github/RWS.

Pixel-space gradient visualizations such as Guided Backpropagation [7] and Deconvolution [8] have been able to discover fine-grained details in the image, but are not class-discriminative and introduce significant processing overhead. Localization approaches like Class Activation Mapping (CAM) [9] and Gradient-weighted CAM (Grad-CAM) [10], are class-discriminative and achieve localization in one shot requiring a single forward and partial backward pass per image. Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision, resulting in heatmaps that are both high-resolution and class-specific. These heatmaps are commonly implemented as attention-based models, which calculate pixel relations and then enhance features through weighted aggregation.

Segmentation methods often incorporate heatmaps to enhance prediction. For example, adding semantic information [11], [12], cumulative activation mapping [13], [14], and chunked segmentation merging [15]. Nevertheless, these strategies are less effective when introduced information contains interfering factors like mutual features between different objects, background, and insufficient heat map resolution which often diminish their effectiveness. Hence, the selection of beneficial features from heatmaps remains a challenge (Fig. 1) as further analysis is required to determine whether the reintroduced information benefits all pixels in low-attention heatmap regions.

RWS introduces a selection strategy to ensure that the reintroduced information benefits all pixels within low-attention regions, ultimately enhancing segmentation quality. Specifically, our method focuses on selecting low-attention regions in heatmaps, i.e., *weak slices*, which have a negative contribution to segmentation accuracy. We observed a positive correlation between the segmentation accuracy (mIoU) of varying-size slices and their average heat map values (see Fig. 2). Consequently, weak slices are identified as those with both low segmentation accuracy (mIoU) and average attention values (calculated using Seg-Grad-CAM) falling below a specified threshold. Our method selects features from these weak slices across different convolutional layers and integrates them back into the network to refine weak slices and enhance segmentation accuracy.

The main contributions of this work are as follows:

- RWS provides a novel feature selection and reintegration strategy that aims to benefit all pixels in low-attention regions and enhance the overall segmentation result.
- RWS is a highly scalable plug-and-play module that can be easily integrated into state-of-the-art CNNs, ensuring a consistent and ongoing enhancement in accuracy.

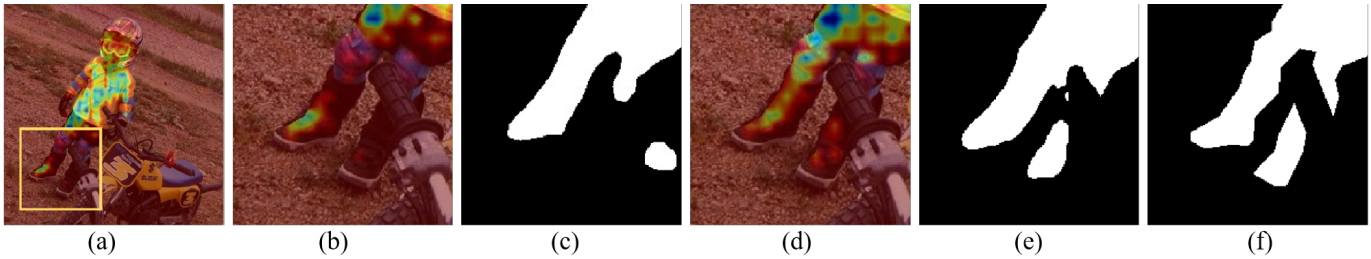


Fig. 1. RWS illustration. Initially, our method identifies regions of low attention in segmentation heatmaps computed by Seg-Grad-CAM (a-b). These regions coincide with low segmentation accuracy (c) and are designated as weak slices. Local learning of weak slices significantly improves their heatmap attention (d), resulting in enhanced segmentation (e) as compared to GT (f).

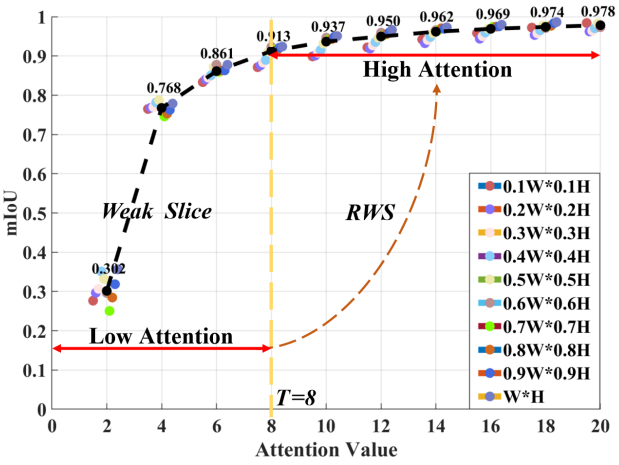


Fig. 2. Weak slice threshold T computation. The graph denotes mIoU vs. slice average attention. The minimum attention value in the graph with a slope less than a given value (0.02 in our experiments) is defined as the weak slice threshold T.

II. Related work

Our related work review primarily focuses on semantic segmentation enhancement methods by investigating the decision-making processes of CNNs and the interpretation of their predictions. These methods specifically explore aspects of feature extraction and their interrelationships in the context of segmentation enhancement.

A. Global Context Learning

To enhance segmentation accuracy, various methods have been proposed to incorporate global and local information sharing [16]. Pyramid Scene Parsing Network (PSPNet) [17] decomposes global contextual information into sub-region contextual information, improving detailed segmentation accuracy. Similarly, Spatial Pyramid-based Graph Reasoning (SpyGR) [18] employs graph convolutional neural networks to capture global motion information and local interactions, thereby improving segmentation accuracy for human actions.

The Context Encoding Network (EncNet) [19] introduces global contextual information to capture the contextual semantics of scenes and selectively highlight feature maps by categories. The Adaptive Context Network (ACNet) [20] measures the similarity between global and local features, resulting in adaptive contextual features. The Asymmetric Non-local Neural Network (ANNet) [21] constructs an adaptive scale interaction network that considers information from other positions, improving the features of each position. The Dynamic Multi-scale Filters Net (DMNet) [22] utilizes context-aware filters to estimate semantic information at specific scales, enabling the segmentation of objects of different scales.

In [23] information from all positions is aggregated, forming bidirectional propagation paths. In [24] expectation-maximization is applied to derive a set of compact bases for attention mechanisms, reducing the complexity of bidirectional information. The Criss-Cross Attention (CCA) module [25] excels in the collection of contextual information by utilizing neighboring pixel information along cross paths. Similarly, authors in [26] share global information through lightweight contextual blocks integrated into each convolutional layer of the backbone network. GCNet [27] enhances per-pixel features by leveraging global context.

The Dual Attention Network (DANet) [28] incorporates a dual-attention network module that considers both local features and their global position dependencies to enhance segmentation accuracy for local details. Disentangled Non-local Neural Network (DNL) [29] leverages self-attention mechanisms to connect local position with global contextual information, enabling to learn long-distance relationships.

A Transformer module that effectively combines low-level and high-level features while maintaining feature consistency was introduced [30]. Global attention multi-resolution Transformers were introduced in [31] to model interactions across all image regions, thus improving the effectiveness of semantic segmentation models. Similarly, Transformer enhanced global features were introduced in [13].

Recently, several works have been focusing on feature leveraging for semantic segmentation enhancement as follows. Feature saliency enhancement through simultaneous feature determination of multiple targets [32]; improvement of perceptual field of DNNs through Background-aware Pooling and Noise-aware Loss (BAP-NAL) in [33]; generation of pseudo-labels to delineate targets and contexts in Beyond Semantic to Instance Segmentation (BESTIE) [34]; additional feature information by considering images from the same class in Cross-Image Affinity Net (CIAN) [35]; high-level contextual features [36].

Nevertheless, these methods require extensive training to converge and achieve satisfactory accuracy. In contrast, we

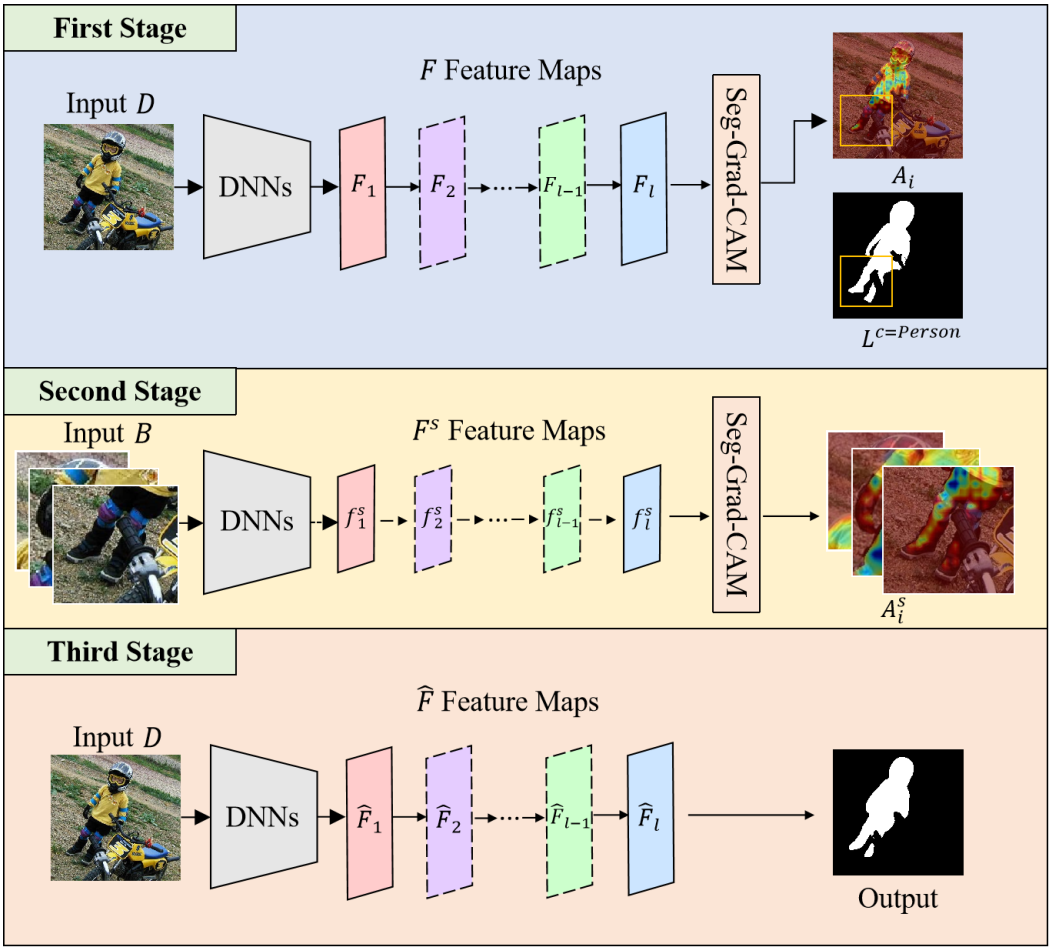


Fig. 3. Overview of RWS 3 stages pipeline. In the first stage (top row), a semantic segmentation DNN on images yields attention maps F_i for each convolutional layer using the Seg-Grad-CAM method. For each segmentation class L^c, we compute weak slices which are low-attention regions conjoining low accuracy segmentation. In the second stage (mid row), we repeat semantic segmentation DNN training on the weak slices obtaining local attention maps per slice F^S. In the third stage (bottom row), we integrate features from different layers in the local slice DNN into the global DNN according to relative attention values in the local and global attention maps. This results in enhanced feature layers $\hat{F}$ and an improved segmentation.

take a CAM approach which achieves prediction in one shot requiring a single forward and partial backward pass per image. We use the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision, resulting in attention heatmaps that are both high-resolution and class-specific.

B. Class Activation Mapping (CAM)

Before the advent of CAM methods, DNN interpretability often relied on analyzing gradient information from feature responses or category activations within convolutional layers. E.g., Gradient-free Activation Maximization (GAM) [37] used generative neural networks and genetic algorithms to interpret DNNs in terms of feature responses.

CAM [10] based methods use gradient information flowing into CNN layers to assign importance values to each neuron, resulting in heatmaps that are both high-resolution and class-specific.

In recent years, CAM has gained widespread attention due to its ability to optimize CAM and its derivative applications [38]–[42]. For example, Score-CAM utilizes activation maps instead of gradients to significantly reduce the computational workload [14]. Ablation-CAM analyzes the importance of each feature map weight to achieve class recognition [43].

Feature map weights were shown significant also for semantic segmentation. Eigen-CAM [44] subtracts complex computational branches from CAM, while Layer-CAM [45] expands the perceptual field for segmentation by introducing fine-grained features.

Our method follows the Seg-Grad-CAM [46], which computes pixel-level focus visualization for segmentation tasks. Our work extends this method by searching weak regions in the resulting heatmaps and enhancing them.

C. Negative Samples Refinement

The training of object detectors often faces the challenge of a substantial class imbalance, with a vast number of background examples compared to the annotated objects of interest. This imbalance has been addressed through techniques like bootstrapping, originally introduced in the mid-1990s, which iteratively grows the set of background examples by selecting those instances where the detector produces false alarms.

Along this line, Online Hard Example Mining (OHEM) [47] is a well-established method that primarily targets challenging negative samples, which are selected according to their low attention and poor accuracy. Unfortunately, this approach neglects a crucial subset of samples—those with low attention but high accuracy which have the potential to improve the overall accuracy. Recognizing this limitation, our RWS technique has been developed to address this gap by reintroducing both types of samples into the training process.

Partial Class Activation Attention (PCAA) [15] suggests splitting the CAM into localized segmentation tasks [48]. Their results demonstrate a greater level of attention at the region level compared to the global solution [38].

Nevertheless, their feature fusion process still faces challenges, including the occurrence of redundant feature overlay and over-enhancement due to duplicate features. Our method solves these problems through the refinement of weak slices and selective integration of features into the DNN.

III. Overview

The RWS method enhances semantic segmentation in three steps (see also Fig. 3):

- **Weak Slices Identification:** In the first step, a DNN is trained for semantic segmentation in conjunction with Seg-Grad-CAM to produce global heatmaps containing per-pixel attention values. The method then identifies regions in the input images where attention values in corresponding heatmaps are lower than a threshold T. These regions are referred to as "weak slices," and they become the focus of enhancement in subsequent steps.
- **Local Features Extraction:** In the second step, our method focuses on these weak slices to learn meaningful local features. To do this, it trains a DNN for slice segmentation coupled with Seg-Grad-CAM, yielding local features, where a DNN for slice segmentation is trained, resulting in local heatmaps.
- **Global Enhancement:** In the third step, a DNN for global segmentation is trained by integrating local features from the previous step with global feature maps. This integration enhances the overall segmentation accuracy by refining the low-attention regions (weak slices) using local features, ultimately improving the quality of semantic segmentation.

IV. Technical Details

As described above, RWS consists of three stages that perform sequentially. In the following, we discuss the technical details of these stages.

A. First Stage: Weak Slices Identification

In this stage, we identify regions with low attention values, namely weak slices. Initially, the input image D is processed by the DNNs, resulting in various convolutional layer features $\{F_i\}$. Each F_i is then fed into the Seg-Grad-CAM algorithm to generate respective heatmaps A_i for each convolutional layer (in Fig. 3, First Stage).

In a nutshell, Seg-Grad-CAM heatmaps [46] are computed as follows:

$$H_i^c = ReLU(\sum_k \alpha_i^k F_i^k), \;\; \alpha_i^k = GAP(\frac{\partial \sum y^c(m,n)}{\partial F_i^k}). \quad (1)$$

$$A_i^c = H_i^c, \;\; if \; logit(y^c) = true. \quad (2)$$

H_i^c represents the heat map results of category c in layer i (i.e., $1, 2, ..., l-1$, l), F_i^k is the feature map of k_{th} kernel of the i_{th} layer, GAP denotes global average pooling, $y^c(m,n)$ is the prediction of pixel (m,n) for class c. We then perform a logit on the prediction of the whole image pixels for a chosen class c to define the attention values A_i^c. α_c^k denotes the weight of the k_{th} convolution kernel. $ReLU$ is used to highlight only the contributing pixels.

Next, we use a sliding window to traverse the image and generate image slices for which we select weak slices w.r.t. class labels. For each slice generated by the sliding window traversal, we compute its average attention value and also its mIoU w.r.t. each segmentation class c.

Weak slice threshold $[T]$. Weak slice selection threshold T is computed by examining the relation between mIoU and average attention values of slices. We analyze the curve representing this relation and define T as the minimum attention value on the curve with a slope below a given value (0.02 in our experiments, see Fig. 2). This means that weak slices are identified as slices that have both low mIoU and low average attention, indicating a high potential for enhancement.

Note that initially, T is computed per class. For computational efficiency, we then average all $T's$ into one. Similarly, T is calculated per dataset and needs to be recalculated for different datasets. Finally, once T is set, all slices below this threshold are selected as weak slices.

B. Second Stage: Local Features Extraction

In the second stage, we compute local features for the weak slices obtained in the first stage. For this purpose, we train our CNN on the weak slice dataset, essentially focusing it on low-attention regions in the input images. Thus, we feed weak slice images and their labels L^{Slice} into the DNN. The result is heat maps A_i^s of each layer (see Fig. 3, Second Stage).

In Fig. 4, we show the effect of attention enhancement by local training on weak slices. Specifically, a weak slice selected from the original input image has low attention values as can be seen in Fig. 4(a,c). Training on weak slices obtains high attention values in meaningful regions as can be seen in Fig. 4(c,d). This yields features in these regions to enhance the overall semantic segmentation.

Nevertheless, not all pixels in the slice's attention layers exhibit higher attention. In fact, during subsequent training iterations, pixels at different layers can diminish the overall enhancement effect. Therefore, we need to carefully select the favorable pixel features across attention layers.

We normalize the pixel attention values across convolutional layers w.r.t. the maximum attention value of a pixel (see also Fig. 5). For a pixel indexed as (m,n), we define:

$$\Gamma(m,n) = max\{A_1(m,n); A_2(m,n); ...; A_l(m,n)\}. \quad (3)$$

(a) A_1 (b) A_1^S (c) A_4 (d) A_4^S

Fig. 4. Slice heatmap enhancement. Figures (a) and (c) are the first and last (correspondingly) heatmap layers obtained from the first stage of global DNN training. Figure (b) and (d) are corresponding heatmap layers obtained from the second stage of local DNN training. Note that the last layer shows significant enhancements in some regions (yellow boxes), but the first layer also reveals that not all pixels in the slices are conducive to increased attention (yellow box). This indicates the necessity of our RWS method for favorable feature selection.

The normalized relative attention of the pixel in layer i is given by $A_i(m,n)/\Gamma(m,n)$. We then define $\omega_i(m,n)$ as the potential of enhancement for the pixel with features obtained from layer i:

$$\omega_i(m,n) = \Gamma(m,n) - A_i(m,n). \tag{4}$$

C. Third Stage: Global Enhancement

In the third stage, we integrate features obtained from the weak slice learning process in the second stage, into the global DNN to achieve semantic segmentation enhancement.

Feature selection $[\omega]$**.** In the integration process, the selection of features is guided by two main principles. First, the pixel's attention and contribution in the slice should be higher than its attention and contribution in the whole image. Second, we utilize only features with high potential for enhancement.

Thus, we evaluate the attention values of per-pixel (m,n) in each slice it appears in and for each layer. The feature of a pixel in slice S and layer i is selected if it simultaneously satisfies the following three conditions:

$$\begin{cases} \omega_i(m,n) > \varepsilon, \ and \\ A_i^s(m,n) > A_i(m,n), \ and \\ A_i^s(m,n)/\Gamma^s(m,n) > A_i(m,n)/\Gamma(m,n). \end{cases} \tag{5}$$

where ε is a small value close to zero, which means the pixel's attention value should be lower than the maximum, indicating the potential for enhancement in the pixel feature. In addition, the other two conditions mean that the selected pixel has a higher attention value in both absolute and relative normalized form.

Feature integration $[\lambda]$**.** Integration of features that fulfill the above condition is defined as:

$$\hat{F}_i(m,n) \leftarrow \sum_i \lambda_i \hat{F}_i(m,n) + (1-\lambda_i) F_i^s(m,n), \tag{6}$$

where the integration weight λ_i is calculated using the ratio of the relative attention value of the pixel in the global and local heatmap:

$$\lambda_i = \frac{A_i(m,n)/\Gamma(m,n)}{A_i(m,n)/\Gamma(m,n) + A_i^s(m,n)/\Gamma^s(m,n)}. \tag{7}$$

During the integration process, the parameter λ is used to calculate the proportion between F and F^s. When the attention of F^s's features is higher, they have a greater influence on $\hat{F}$. On the other hand, when the attention of F is higher, $\hat{F}$ aims to preserve the feature values of F. This approach updates features of $\hat{F}$ adaptively and enhances the attention of local pixels across different convolutional layers, ultimately improving segmentation accuracy. Finally, RWS provides a DNN that integrates selective features from across different layers for enhanced semantic segmentation.

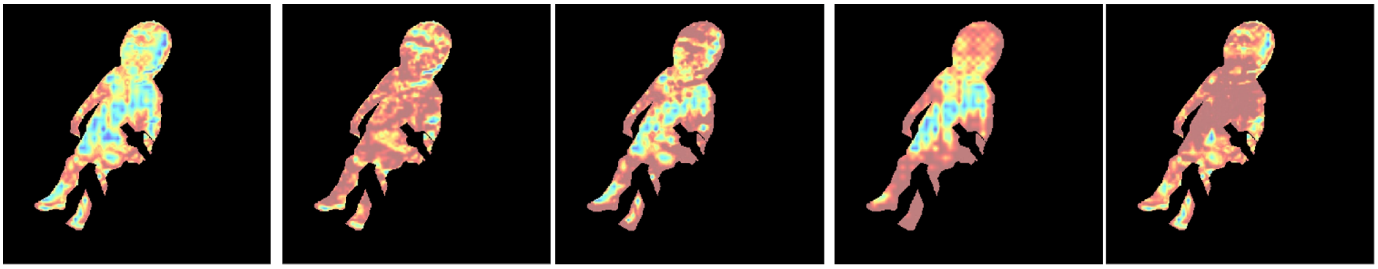
It is possible to repeat the three stages process iteratively until the number of weak slices does not decrease and stabilizes. In this case, no further enhancement is obtained. Thus, in the case of repeated RWS iterations, when the number of weak feature slices should stabilize, indicating convergence, RWS becomes ineffective in further enhancing the segmentation accuracy.

V. Results

Datasets. Experiments are conducted on four widely used benchmarks of semantic segmentation.

- **VOC2012** [49] dataset is a semantic segmentation dataset consisting of 20 object classes and a total of 13,487 images, divided into 10,582 training, 1,449 validation images, and 1,456 testing images. VOC2012 enhanced datasets are employed as test cases.
- **COCOStuff** [50] is a scene segmentation dataset that includes 182 object annotations and 80 scene annotations. It consists of 80,000 training, 5,000 validation, and 5,000 testing images.
- **ADE20K** [51] is a scene segmentation dataset containing 25K images with 20K different categories. The dataset is divided into 20k training, 2k validation, and 3k testing images.
- **Cityscapes** [52] is a segmented dataset of urban scenes, featuring 19 semantic categories. It includes 5,000 high-quality pixel-level annotated images and 20,000 coarse annotated images. The 5,000 finely annotated images are further split into 2,975 training images, 500 validation images, and 1,525 testing images.

Implementation Details. Our backbone is initialized with pre-trained weights on ImageNet. During the training process, batch processing normalization was performed with PyTorch.



(a) Γ (b) A_1 (c) ω_1 (d) A_4 (e) ω_4

Fig. 5. Γ, A_i, ω_i examples. Left-to-right is Γ heatmap (a) for segment label human. A_i heat maps next to corresponding ω_i's for layers 1 and 4 respectively.

Training was conducted on four NVIDIA GeForce RTX 3090 GPUs with 24 GB memory per card, and testing was performed on a single NVIDIA GeForce RTX 3090.

Currently, RWS processes images with a maximum resolution of 5368 $\times$ 4026 pixels. Handling higher-resolution images may exceed memory limits. However, the computation of features and weights for each image can be independently exported at each stage and can be stored on a hard drive. These features and weights are read separately when needed in the third stage.

Effectively there are three training stages in RWS: F, F^s, $\lambda\hat{F}$ + $(1-\lambda)F^s$. Training settings for different datasets are as follows:

- **VOC2012**: Initial learning rate is 0.01, weight decay is 0.0005. Images are uniformly cropped to a size of 512 $\times$ 512, with batch size set to 16 by default. The model is trained for 30K iterations.
- **COCOStuff**: Initial learning rate is 0.001, weight decay is 0.0001. Images are uniformly cropped to a size of 512 $\times$ 512, with batch size set to 16 by default. The model is trained for 60K iterations.
- **ADE20K**: Initial learning rate is 0.01, weight decay is 0.0005. Images are resized to 512 $\times$ 512, and the batch size is set to 16 by default. If unspecified, the model is fine-tuned for 160K iterations.
- **Cityscapes**: Initial learning rate is 0.01, weight decay is 0.0005. Images are uniformly cropped to a size of 512 $\times$ 1024, and batch size is set to 8 by default. The model is trained for 80K iterations if no other number is specified.

TABLE I
ABLATION STUDY RESULTS ON DEEPLAB v_3 WITH RESNET-101 BACKBONE W.R.T. WEAK SLICES AND $[\omega]$ ON VOC2012 DATASET.

	Arbitrary Slices	Weak Slices	
		no $[\omega]$	$[\omega]$
mIoU	85.9	86.3	87.3

A. Results

Our Refined Weak Slices method focuses on learning local features in weak slices and then integrating them into the global DNN enhancing segmentation results. In Fig. 1 we demonstrate our segmentation enhancement. Initially, the left leg has low segmentation accuracy (c) compared to GT (f).

TABLE II
ABLATION STUDY RESULTS ON COCOSTUFF DATASET AND RESNET-50 BACKBONE W.R.T, WEAK SLICES, $[\omega]$ AND $[\lambda]$.

Baseline	Weak Slice	$[\omega]$	$[\lambda]$	mIoU
✓				40.1
✓	✓			42.2
✓	✓	✓		42.9
✓	✓		✓	42.7
✓	✓	✓	✓	43.9

Our method identifies low-attention regions in the Seg-Gtad-CAM heatmap (a), defining weak slices (b) in conjunction with low segmentation accuracy (c). Locally training on weak slices yields features that are integrated in the global segmentation and enhance segmentation accuracy in these regions (e).

Similarly, Fig. 4 demonstrates the attention enhancement achieved in the local weak slice learning stage of our method. We observe that by focusing on low attention weak slices (a,c) allows learning meaningful local features which are then integrated to enhance overall segmentation (b,d).

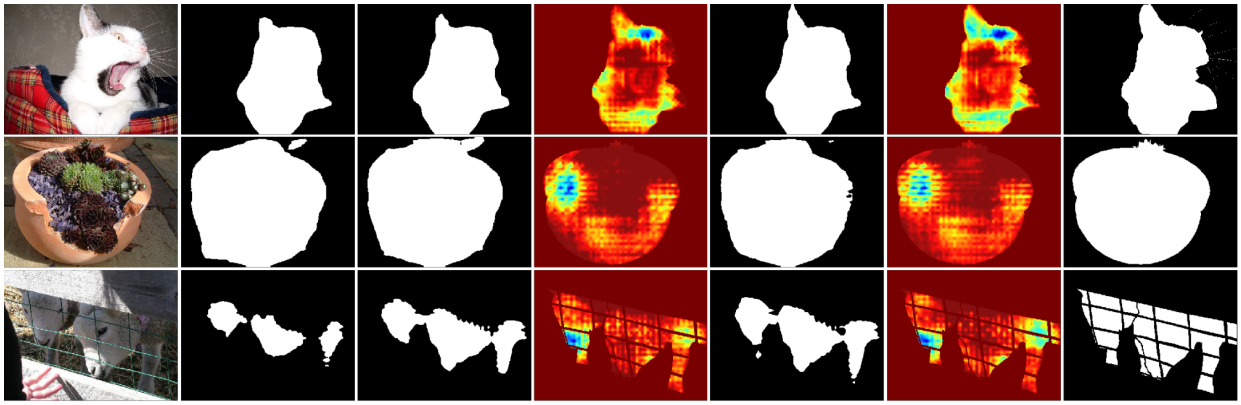
B. Ablation Study

We perform an ablation study to evaluate the benefit of weak slices for semantic segmentation enhancement.

Weak Slices contribution. Table I demonstrates segmentation accuracy of DeepLab v_3 with ResNet-101 on VOC2012 dataset. We compare arbitrary slices (left col.) vs. weak slices (mid, right cols.). Among weak slices, we evaluate our feature selection policy by comparing no feature selection (no $[\omega]$, i.e., integrate all features in weak slices) vs. our conditional feature integration ($[\omega]$). We observe that weak slices and their refinement achieve the highest segmentation mIoU accuracy.

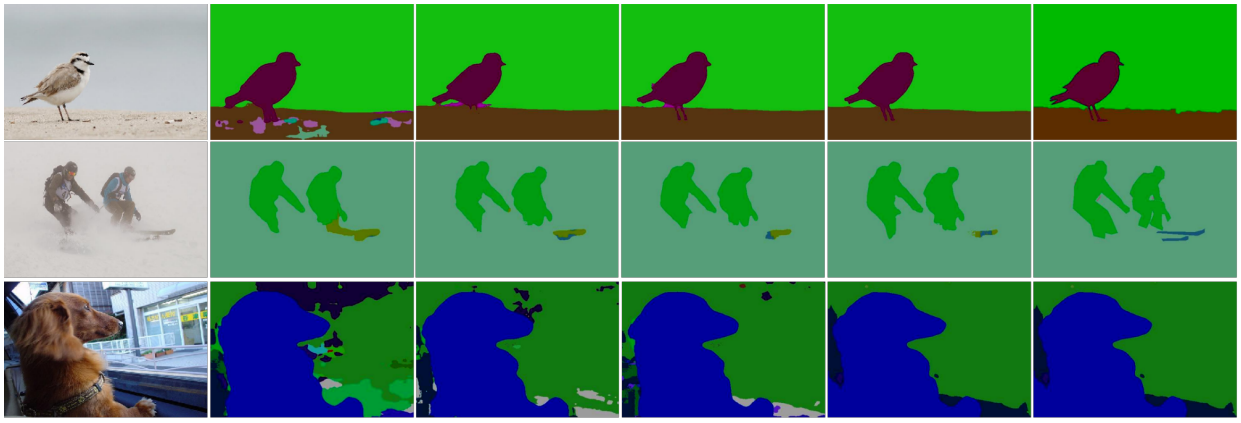
Fig. 6 shows qualitative segmentation results for the weak slice ablation study with DeepLab v_3 and ResNet-101 on VOC2012 dataset. It is evident that our weak slice refinement enhances segmentation accuracy when compared to baseline and arbitrary slices. Additionally, arbitrary slices with no selective feature integration tend to enlarge segmentation areas and introduce errors (see mid-row). RWS outperforms both arbitrary slice enhancement and baseline.

$[\omega]$, $[\lambda]$ **parameters effect.** We perform an ablation study to validate the algorithmic choices of our RWS method on the



(a) Image (b) Baseline (c) Arbitrary Slices (d) Heat-Slices (e) RWS (f) Heat-RWS (g) GT

Fig. 6. Ablation study qualitative results on VOC2012 dataset with DeepLab v_3 and ResNet-101. Left-to-right, initial image (a), no slice segmentation (b), arbitrary slice enhancement (c), weak slice refinement (d), and GT (e).



(a) Image (b) Baseline (c) Weak Slices (d) ω (e) RWS (f) GT

Fig. 7. Ablation study qualitative results on COCOStuff with ResNet-50 w.r.t. weak slices, $[\omega]$, and $[\lambda]$.

COCOStuff dataset. Specifically, we explore the effectiveness of weak slices, $[\omega]$, and $[\lambda]$.

Table II summarises and compares these results. The rows show segmentation accuracy (mIoU) using only ResNet-50 as the backbone and FCN [53] (top), integrating weak slices as a whole, without feature selection, integrating conditioned features $[\omega]$, and using weighted feature integration $[\lambda]$. We observe that the full RWS algorithm achieves the highest accuracy.

Similarly, Fig. 7 presents three qualitative results of the ablation study. While weak slices already show an improvement over FCN, feature selection $[\omega]$ and weighted feature integration $[\lambda]$ provide the most accurate segmentation results. As shown in Fig. 8, (c) is the segmentation result after feature enhancement in the whole weak slice including both inside and outside of the class c region, which may lead to incorrect segmentation. A_i^c restricts the feature selection within the class region to minimize interference from outside pixels.

$\hat{\mathbf{F}}_\mathbf{i}$ layers contribution. Next, we evaluate the contribution of different convolutional layers in the segmentation enhancement. Table III summarises the segmentation accuracy (mIoU) for the utilization of specific layers in the RWS feature computation and integration.

(a) H_i^c (b) Weak Slice (c) Output (d) GT

Fig. 8. Ablation study of A_i^c. Feature enhancement in the whole weak slice, especially outside the target region, may lead to incorrect segmentation (c).

We observe that not using any feature enhancement results in the lowest mIoU. Furthermore, it is noted that early layers $(1, 2)$ contribute less to enhancement compared to later layers $(3, 4)$. Nevertheless, it is noticeable that all layers have a unique contribution to segmentation accuracy as there is a consistent increase in mIoU score with the addition of layers.

T threshold effect. T is the threshold defining weak slices, specifically denoting the slice average attention to mIoU relation. The larger T the more slices are considered as weak slices and inversely. Fig. 9 depicts the segmentation mIoU w.r.t. different T values. Naturally, increasing T leads to more slices and our DNN learns local features to enhance them, thus improving the mIoU (at a performance cost).

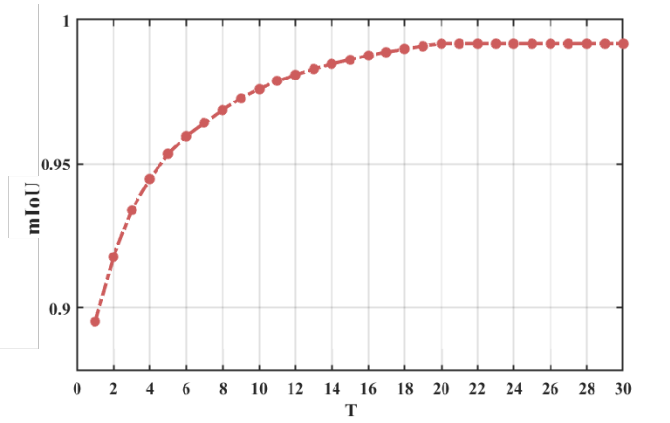


Fig. 9. Ablation on weak slice selection threshold T.

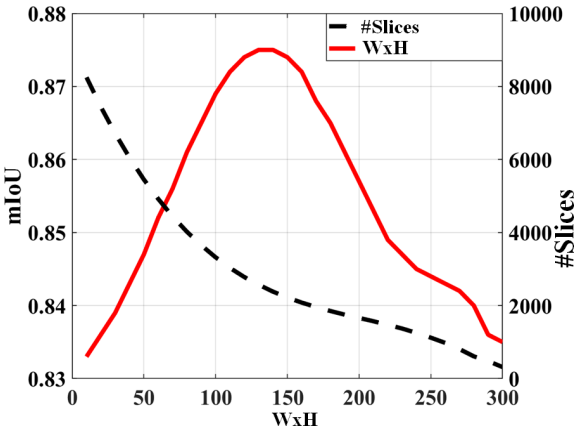


Fig. 10. Ablation study of $W \times H$ size vs. number of slices and mIoU on DeepLab v_3 with ResNet-101.

TABLE III
ABLATION STUDY ON USING DIFFERENT CONVOLUTIONAL LAYERS IN THE FEATURE ENHANCEMENT AND INTEGRATION WITH DEEPLAB v_3 AND RESNET-50 AS THE BACKBONE TESTED ON THE VOC2012 TEST SET.

$\hat{F}_1$	$\hat{F}_2$	$\hat{F}_3$	$\hat{F}_4$	mIoU	$\hat{F}_1$	$\hat{F}_2$	$\hat{F}_3$	$\hat{F}_4$	mIoU
-	-	-	-	78.9	-	✓	✓	-	82.3
✓	-	-	-	80.6	-	✓	-	✓	82.4
-	✓	-	-	81.2	-	-	✓	✓	81.8
-	-	✓	-	80.8	-	✓	✓	✓	84.0
-	-	-	✓	81.4	✓	-	✓	✓	83.2
✓	✓	-	-	81.8	✓	✓	-	✓	83.2
✓	-	✓	-	81.9	✓	✓	✓	-	83.5
✓	-	-	✓	82.2	✓	✓	✓	✓	84.2

W×H effect. Fig. 10 shows the impact of weak slice size on the number of slices and segmentation accuracy (mIoU). Obviously, as the weak slice size ($W \times H$) decreases, the number of weak slices increases (dotted curve) which influences training times.

As can be seen, in the case of too small or too large slices, the average attention vs. mIoU is too noisy and does not accurately yield true weak slices. In both extremes, we observe that mIoU is low in comparison with medium-size slices. Thus, we found that slices of 128×128 efficiently capture local low attention and yield significant features to enhance segmentation.

Repeated RWS iterations. We demonstrate the effect of repeated RWS iterations in Fig. 11. Four different examples are presented, each undergoing one, two, three, and four RWS iterations. We show for each example row pairs of the last layer attention map (top) and resulting segmentation (bottom) per iteration.

It is observed that with each repetition, certain local attention values increase, resulting in more refined local segmentation. Meanwhile, other attention values decrease as corresponding local segmentation accuracy stabilizes.

C. Comparison with State-of-the-Art

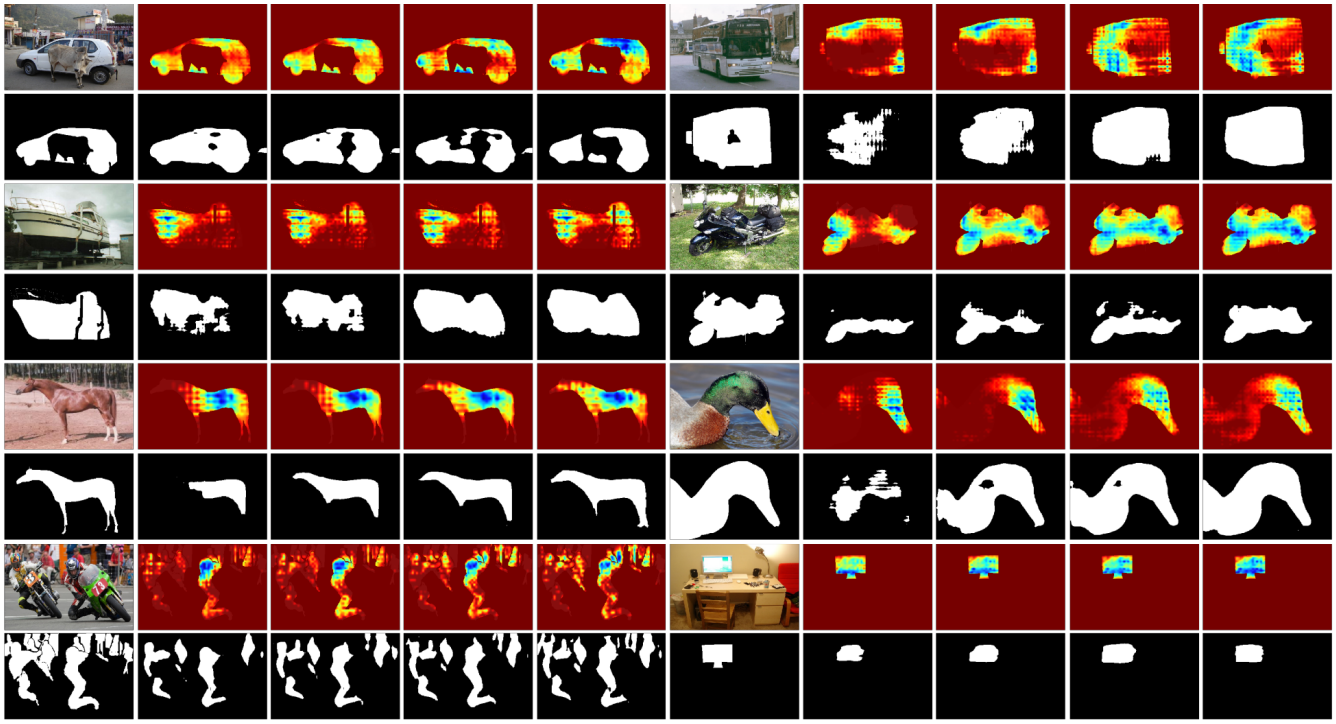
We compare RWS with state-of-the-art segmentation methods. Fig. 12 illustrates qualitative comparisons between our RWS enhancement and three State-of-the-Art (SOTA) methods using the COCOStuff dataset. Across eight input images representing various classes (column (a)), the visual results showcase significant improvements in original segmentation (left columns in (c), (d), and (e)) due to our DNN+RWS enhancement (right columns in (c), (d), and (e)). The effectiveness of this enhancement is attributed to RWS's capability to aid DNNs in refining segmentation accuracy within weak feature regions.

In Tables IV and V, we summarize quantitatively these experiments, demonstrating the enhancement of segmentation accuracy (i.e., mIoU) by adding our weak slice refinement to different DNNs (DNN+RWS). In all experiments, we have run DNNs with the same amount of training epochs as our RWS method.

Table IV compares with state-of-the-art methods on the VOC2012 and COCOStuff datasets. Table V compares ADE20K and Cityscapes datasets. All tables show for each method, mIoU results for both ResNet-101 and ResNet-50 backbones compared to enhancement with our RWS method. Note that all experiments unanimously indicate that our weak slice refinement enhances segmentation and yields higher accuracy for all state-of-the-art methods. Similarly, in both ResNet-50 and ResNet-101 backbones, RWS has provided significant enhancement. In many cases, we observe that RWS improves the segmentation results of ResNet-50-based methods to the level of ResNet-101. Thus, RWS may be utilized as a valuable tool to compensate for the weak learning ability of compact backbones.

In Table IV(VOC2012) we observe that RWS+PCAA on ResNet-101 achieves the highest result 89.2 mIoU and has average improvement of 1.87, and maximum improvement of 3.7 in mIoU over non-RWS methods. This emphasizes the beneficial impact of enhancing attention on weak slice regions for improving the accuracy of DNNs.

Table IV(COCOStuff) compares the performance of RWS to eight state-of-the-art methods on the COCOStuff dataset. This dataset poses significant challenges due to its diverse semantic categories. Nevertheless, RWS+DNN surpasses all other methods on both ResNet-101 and ResNet-50 backbones,



(a) Image/GT (b) 1st iter (c) 2nd iter (d) 3rd iter (e) 4th iter (f) Image/GT (g) 1st iter (h) 2nd iter (i) 3rd iter (j) 4th iter

Fig. 11. RWS iterations (1-to-4) on 4 examples. For each row pair, the top is the image and last layer heatmaps, bottom is GT and resulting segmentations.

achieving an average of 2.98, and maximum of 3.8 mIoU improvement. RWS+HyperSeg achieves the highest accuracy on the COCOStuff test set, with a mIoU of 47.7. These results highlight the effectiveness of RWS in enhancing segmentation performance on both ResNet-101 and ResNet-50 backbones.

Table V(ADE20K) compares RWS to seven state-of-the-art methods on the challenging ADE20K dataset. This dataset is known for its diverse image scales and abundant semantic classes, which pose significant challenges to segmentation algorithms.

Also here RWS demonstrates improvements in accuracy across all state-of-the-art methods, utilizing both ResNet101 and ResNet50 as backbone networks. On average, RWS achieves an overall mIoU improvement of 2.68, and maximum improvement of 3.7. RWS+BAP-NAL achieves the highest accuracy with a mIoU of 49.9. These results demonstrate the ability of our method to detect and enhance features at different scales.

Table V(Cityscapes) compares RWS with eight state-of-the-art methods on the Cityscapes dataset. This table also demonstrates that RWS further enhances the performance of all DNNs, on both ResNet-50 and ResNet-101 backbones. We can see that the improvement achieved by RWS+DNN on ResNet-50 is comparable to that on ResNet-101. On average, RWS achieves an overall mIoU improvement of 3.87, and maximum improvement of 5.8. The highest mIoU of 86.1 is achieved by RWS+DNL.

TABLE VI
COMPARISON BETWEEN TRANSFORMERS AND OPTIMAL DNN+RWS CONFIGURATIONS ON FOUR DATASETS.

Method	VOC.	COCO.	ADE.	City.
ViT-B	87.49	47.37	48.06	84.28
Swin-S	88.76	47.61	49.37	86.05
DNN	86.54	44.82	47.13	81.98
DNN+RWS	89.21	47.73	49.89	86.14
	(PCAA)	(HyperSeg)	(BAP-NAL)	(DNL)

In table VI we also compare our DNN+RWS enhancement with Transformers. We compare ViT-B with a batch size of 16 [55] and Swin-S [56] on our four test sets vs. the best performing CNN+RWS configurations. As can be seen in the table, the optimal combination of RWS and CNNs still outperforms Transformers.

VI. CONCLUSIONS AND LIMITATIONS

We propose RWS, a method to enhance the segmentation accuracy of arbitrary DNNs with convolutional layers. Our method defines weak slices as low-attention regions in different convolutional layers which also yield low segmentation accuracy. RWS locally trains on weak slices and extracts features that are then integrated into the global DNN to achieve higher segmentation accuracy. The effectiveness and generalizability of RWS are validated through experiments conducted on four segmentation datasets. We also compare RWS with several SOTA segmentation DNNs to demonstrate the advantages of our technique.

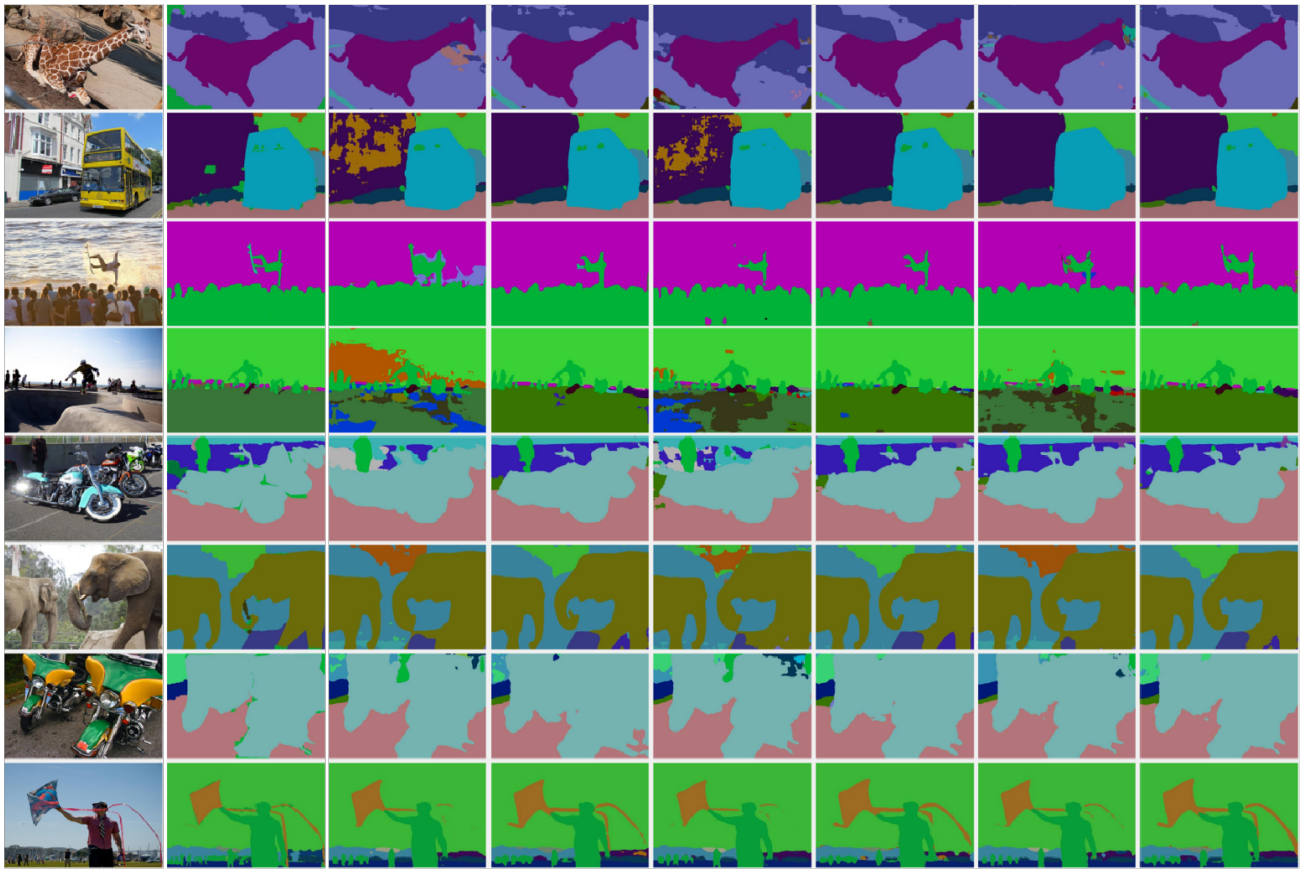


Fig. 12. Qualitative comparisons (rows) of our RWS enhancement with SOTA (ResNet-50 backbone) on COCOStuff dataset. Left-to-right, input image (a), GT (b) segmentation labels, ACNet [20] vs. ours (c), CIAN [35] vs. ours (d) and HyperSeg [36] vs. ours (e).

Fig. 13. Incorrect weak slices w.r.t. class label.

In the future, the RWS concept can be investigated also in the context of non-convolutional networks, such as Transformers, which are currently showing promising results in various tasks. An interesting direction is to accommodate weak slice enhancement to Transformer architectures. Another direction is to adapt RWS to a larger variety of learning tasks, specifically regression and generative models such as GANs and Diffusion Models. It is still unclear how RWS enhancement can be integrated into such pipelines.

Limitations. In terms of limitation, we observe that RWS demonstrates less evident improvements in scenarios involving target-background confusion, extremely small targets, or targets positioned at the edge of the slice (see Fig. 13). In such cases, weak slices may be obsolete and without significant improvement to the overall result.

From a performance standpoint, RWS involves three training stages which introduce an overhead when compared with standard DNNs. Nevertheless, RWS focuses on low-attention regions and enhances them in the overall result. It is possible to consider in future work, ways to shorten global segmentation training stages which are compensated by local weak slice training stages achieving similar or even higher accuracy.

REFERENCES

[1] 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 (CVPR)*, July 2017.

[2] 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.

[3] Y Yuan, L Huang, J Guo, C Zhang, X Chen, and J Wang. Ocnet: Object context network for scene parsing. arxiv 2018. *arXiv preprint arXiv:1809.00916*, 2018.

[4] 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.

[5] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[6] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 593–602, 2019.

TABLE IV
COMPARISONS WITH SOTA ON VOC2012 AND COCOSTUFF DATASETS. AS WITH THE RWS, EACH METHOD RUNS DNN TRAINING FOR 3 EPOCHS.

VOC2012			
Method	Backbone	mIoU	RWS
Affinity [13]	ResNet-101	86.5	88.1
ICCV'2021 (*our impl.*)	ResNet-50	83.8	85.4
ANNet [21]	ResNet-101	84.5	86.2
ICCV'2019 (*our impl.*)	ResNet-50	84.2	85.5
BAP-NAL [33]	ResNet-101	85.3	86.5
ICCV'2021	ResNet-50	83.9	85.6
BESTIE [34]	ResNet-101	85.1	86.6
ICCV'2021	ResNet-50	84.0	85.4
DANet [28]	ResNet-101	84.3	86.2
ICCV'2019	ResNet-50	83.6	85.7
DeepLab v_3 [54]	ResNet-101	85.7	87.3
CVPR'2017	ResNet-50	82.9	84.3
HyperSeg [36]	ResNet-101	84.8	86.5
ICCV'2021	ResNet-50	83.4	<u>87.1</u>
PCAA [15]	ResNet-101	86.5	**89.2**
ICCV'2021 (*our impl.*)	ResNet-50	83.4	86.2
COCOStuff			
Method	Backbone	mIoU	RWS
ACNet [20]	ResNet-101	40.1	<u>43.9</u>
ICCV'2019	ResNet-50	38.3	40.9
CIAN [35]	ResNet-101	43.2	46.1
AAAI'2020 (*our impl.*)	ResNet-50	40.7	43.3
DANet [28]	ResNet-101	39.7	43.0
ICCV'2019	ResNet-50	37.2	40.0
DeepLab v_3 [54]	ResNet-101	40.7	42.9
CVPR'2017 (*our impl.*)	ResNet-50	38.5	40.7
EMANet [24]	ResNet-101	39.9	42.1
ICCV'2019	ResNet-50	37.0	40.5
HyperSeg [36]	ResNet-101	44.8	**47.7**
ICCV'2021 (*our impl.*)	ResNet-50	43.1	45.8
JLSD [32]	ResNet-101	42.5	45.2
ICCV'2019 (*our impl.*)	ResNet-50	41.1	43.7
SpyGR [18]	ResNet-101	39.9	42.9
ICCV'2019	ResNet-50	38.1	41.2
SVCNet [26]	ResNet-101	39.6	43.4
ICCV'2019	ResNet-50	37.2	40.7

TABLE V
COMPARISONS WITH SOTA ON ADE20K AND CITYSCAPES DATASETS. AS WITH THE RWS, EACH METHOD RUNS DNN TRAINING FOR 3 EPOCHS.

ADE20K			
Method	Backbone	mIoU	RWS
ACNet [20]	ResNet-101	45.9	48.9
ICCV'2019	ResNet-50	43.0	46.0
ANNet [21]	ResNet-101	45.2	<u>48.9</u>
ICCV'2019	ResNet-50	42.9	45.7
BAP-NAL [33]	ResNet-101	47.1	**49.9**
ICCV'2021 (*our impl.*)	ResNet-50	44.9	47.8
BESTIE [34]	ResNet-101	47.3	49.1
ICCV'2021 (*our impl.*)	ResNet-50	45.5	48.4
CCNet [25]	ResNet-101	45.2	48.9
ICCV'2019	ResNet-50	43.1	45.7
DANet [28]	ResNet-101	46.2	47.2
ICCV'2019 (*our impl.*)	ResNet-50	43.3	44.8
DMNet [22]	ResNet-101	45.5	48.7
ICCV'2019	ResNet-50	42.6	45.2
Cityscapes			
Method	Backbone	mIoU	RWS
ANNet [21]	ResNet-101	81.3	83.1
ICCV'2019	ResNet-50	77.8	79.6
CCNet [25]	ResNet-101	79.8	84.2
ICCV'2019	ResNet-50	78.5	80.9
DANet [28]	ResNet-101	81.5	85.1
ICCV'2019	ResNet-50	77.2	82.2
DNL [29]	ResNet-101	82.0	**86.1**
ECCV'2020	ResNet-50	79.8	82.9
EncNet [19]	ResNet-101	78.6	82.2
CVPR'2018 (*our impl.*)	ResNet-50	75.2	<u>81.0</u>
GCNet [27]	ResNet-101	79.0	83.1
ICCV'2019 (*our impl.*)	ResNet-50	76.7	82.2
PSANet [23]	ResNet-101	77.9	82.9
ECCV'2018	ResNet-50	76.7	80.1
PSPNet [17]	ResNet-101	80.2	83.7
ICCV'2017	ResNet-50	77.3	82.1

[7] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[8] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[11] Omkar Thawakar, Sanath Narayan, Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Muhammad Haris Khan, Salman Khan, Michael Felsberg, and Fahad Shahbaz Khan. Video instance segmentation via multi-scale spatio-temporal split attention transformer. *arXiv preprint arXiv:2203.13253*, 2022.

[12] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022.

[13] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation

with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022.

[14] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[15] Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, and Qi Tian. Partial class activation attention for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16836–16845, 2022.

[16] 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.

[17] 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.

[18] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020.

[19] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[20] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6757, 2019.

[21] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019.

[22] 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.

[23] 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.

[24] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019.

[25] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[26] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019.

[27] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[28] 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.

[29] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 191–207. Springer, 2020.

[30] Xiang Pan and Jiapeng Xiong. Dctnet: A hybrid model of cnn and dilated contextual transformer for medical image segmentation. In *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 6, pages 1316–1320. IEEE, 2023.

[31] Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, and Alexandre Hostettler. Full contextual attention for multi-resolution transformers in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3224–3233, 2023.

[32] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7223–7233, 2019.

[33] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021.

[34] Beomyoung Kim, Youngjoon Yoo, Chae Eun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4278–4287, 2022.

[35] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020.

[36] Yuval Nirkin, Lior Wolf, and Tal Hassner. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4061–4070, 2021.

[37] Will Xiao and Gabriel Kreiman. Gradient-free activation maximization for identifying effective stimuli. *arXiv preprint arXiv:1905.00378*, 2019.

[38] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022.

[39] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv e-prints*, pages arXiv–2011, 2020.

[40] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[41] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[42] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2117–2125, 2022.

[43] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[44] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[45] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[46] Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13943–13944, 2020.

[47] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[48] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022.

[49] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[50] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[52] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and

Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[53] 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.

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

[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[56] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Yunbo Rao is an Associate Professor at the University of Electronic Science and Technology of China. He received the B.S. degree from Sichuan Normal University, Chengdu, China, in 2003, and the M.E. and Ph.D. degrees from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, in 2006 and 2012, respectively. His research interests include image segmentation, 3-D reconstruction, video enhancement, and medical image processing.

Qingsong Lv received his Master's degree from the School of Information and Software Engineering, University of Electronic Science and Technology of China. His research interests include interpretable neural networks, semantic segmentation, medical image processing and industrial vision.

Andrei Sharf is an Associate Professor at the Computer Science Department, Ben-Gurion University. He received his Ph.D. degree in the School of Computer Science, Tel-Aviv University. His main research interests include geometry processing, 3D modeling, interactive techniques, and deep learning techniques for 3D data.

Zhanglin Cheng is a Professor at the Shenzhen Key Laboratory of Visual Computing and Analytics (VisuCA), Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. He received the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2008. His research interests include computer vision, computer graphics, and visualization.