Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers-CVPR'21

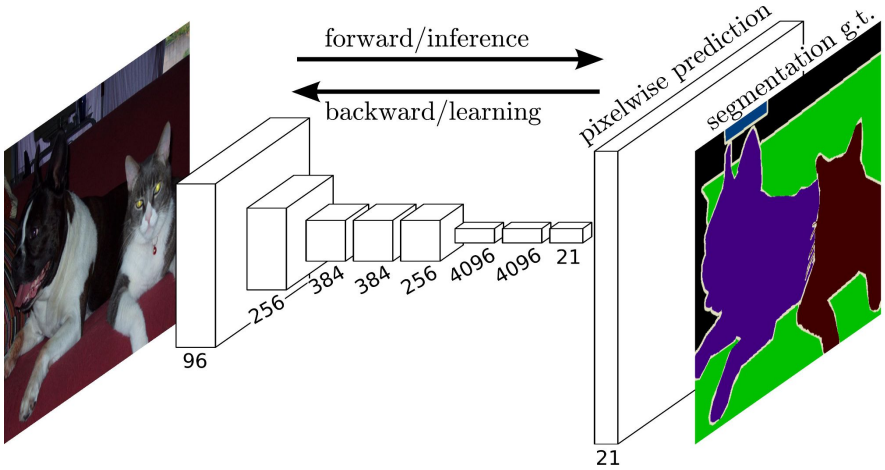
Alejandro Zamora, Brandon Clark, Joshua Gracie, Dunquan Zheng

Introduction/Motivation

- 444 Citations
- Most semantic segmentation methods adopt FCN (fully-convolutional network)
- Aim to provide an alternative perspective for semantic segmentation
  - Sequence-to-sequence prediction
  - Pure transformer Encoder (ViT)
    - Without convolution
    - Resolution reduction

Background

- Semantic segmentation
  - FCN



Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

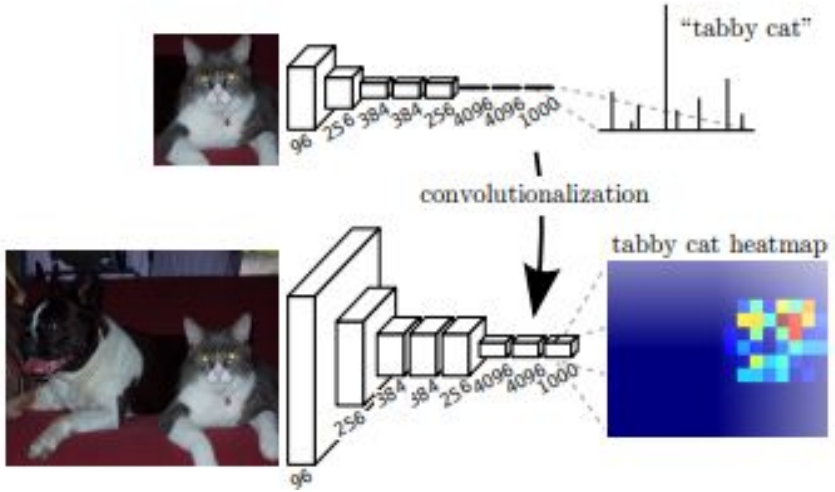
- Transformer
  - DETR
  - STTR and LSTR
  - ViT
  - Axial (Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation)

Axial VS. SETR

- Axial
  - conventional FCN design in spatial resolution of feature maps
    - reduced progressively
  - adopts a specially designed axial-attention
    - less scalable
- SETR
  - sequence-to-sequence prediction model
    - same spatial resolution
  - standard self-attention design
    - maximize the scalability
    - Easy-to-use
  - superior in segmentation accuracy

FCN Segmentation

- Convolutional
  - Replace fully connected layer with corresponding convolutional layer
- Upsample(deconvolution)
  - Recovering position feature of image
- Skip layer
  - Connect output of convolutional layer to deconvolution layer

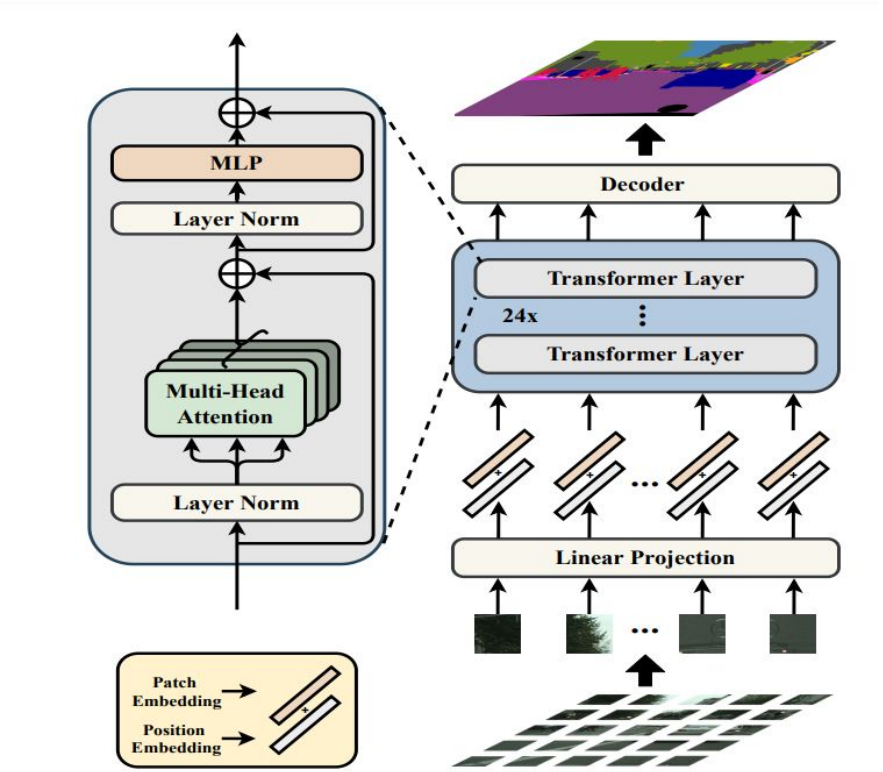


Methods: Encoder

Standard transformer encoder as seen in ViT and others

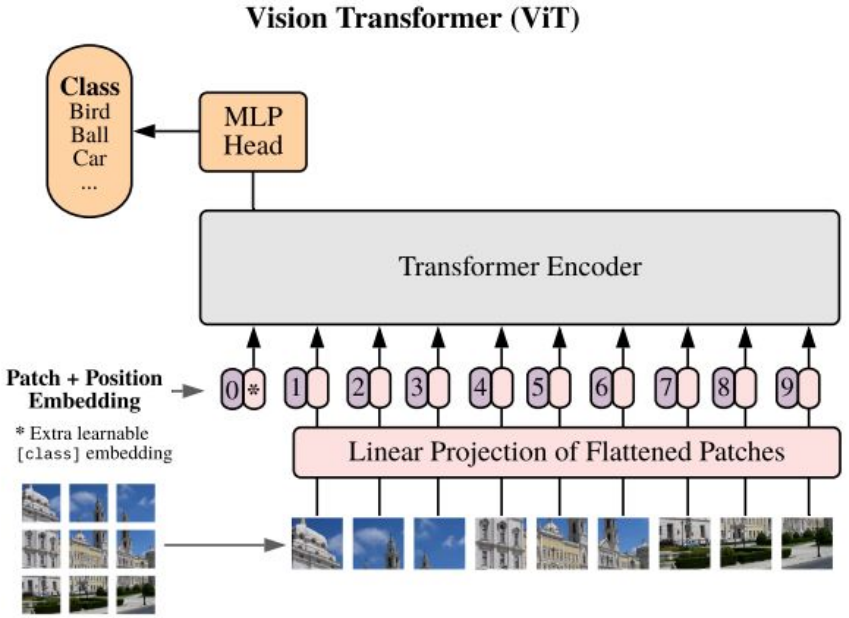
Authors utilize a 16x16 patch size across all ablations

Difference is in the decoder



Encoder Variants

- T-Base
  - 12 Transformer layers
- T-Large
  - 24 Transformer layers
- Hybrid Baseline
  - ResNet-50 based FCN Encoder
  - Output fed to SETR Decoder
- Pre-Training
  - ViT or DeiT on ImageNet 1k/21K



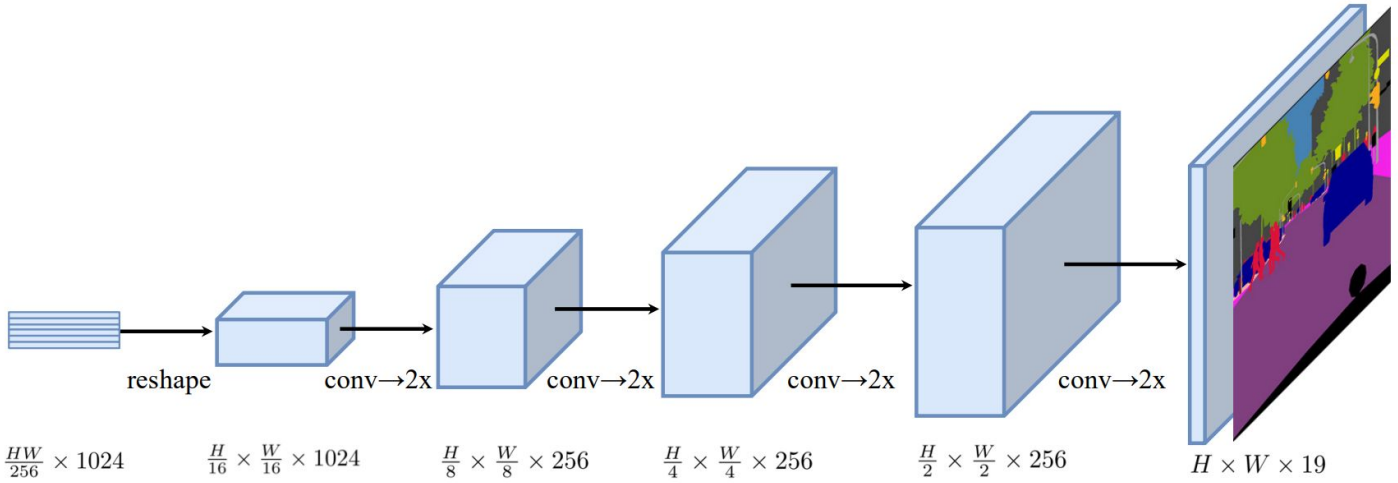
Methods: Naive Upsampling

2 layer upsampling method:

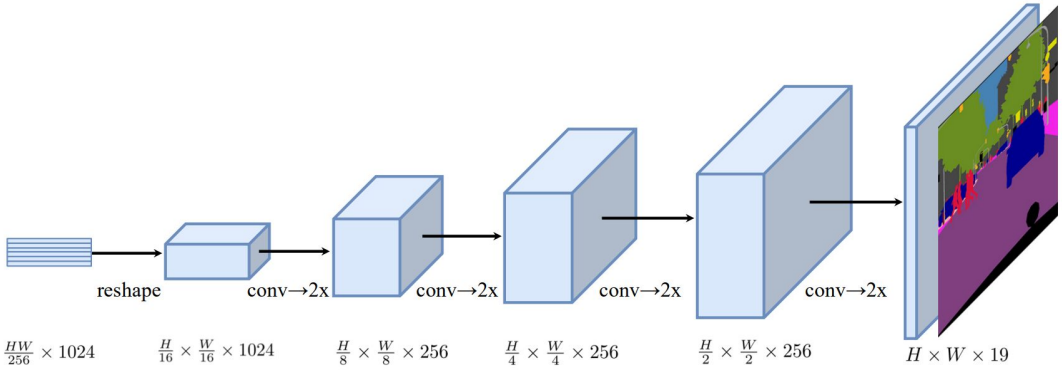
1. 1x1 conv -> batch norm (w/ ReLU)
2. 1x1 conv
3. Bilinearly upsampled to original image size

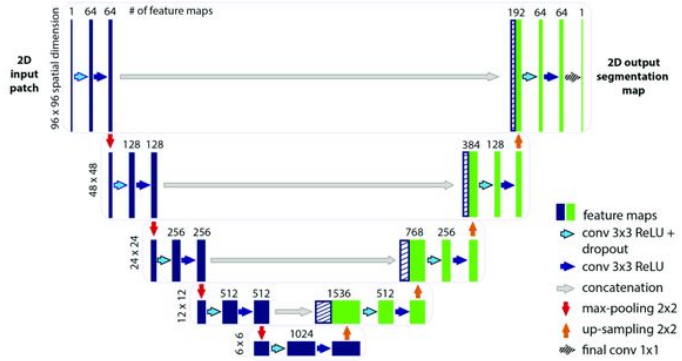
Methods: Progressive Upsampling

- Instead of upsampling at the end, alternate 2x upsampling and conv layers
- Do this operation 4x to get original size image



UNet and PUp





Livne, Michelle, et al. "A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease." *Frontiers in neuroscience* (2019): 97.

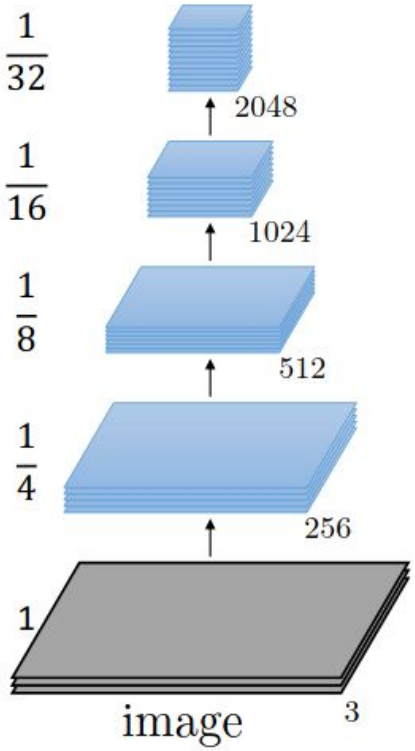
Methods: Multi-Level Feature Aggregation

Similar to multi-level pyramid

Every layer of a transformer outputs same size output

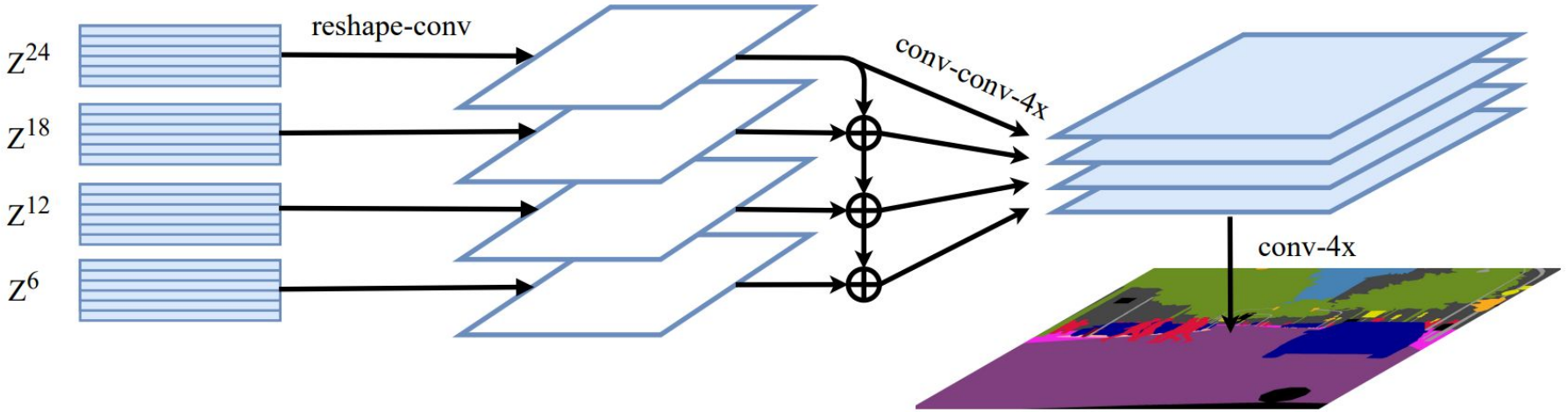
Output features from each layer:

- Reshape output for convolution
- Element-wise addition across layers
- Pass to 3 layer conv -> upscale 4x
- Fuse features along the channel
- Conv + upscale 4x



http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf

Methods: Multi-Level Feature Aggregation

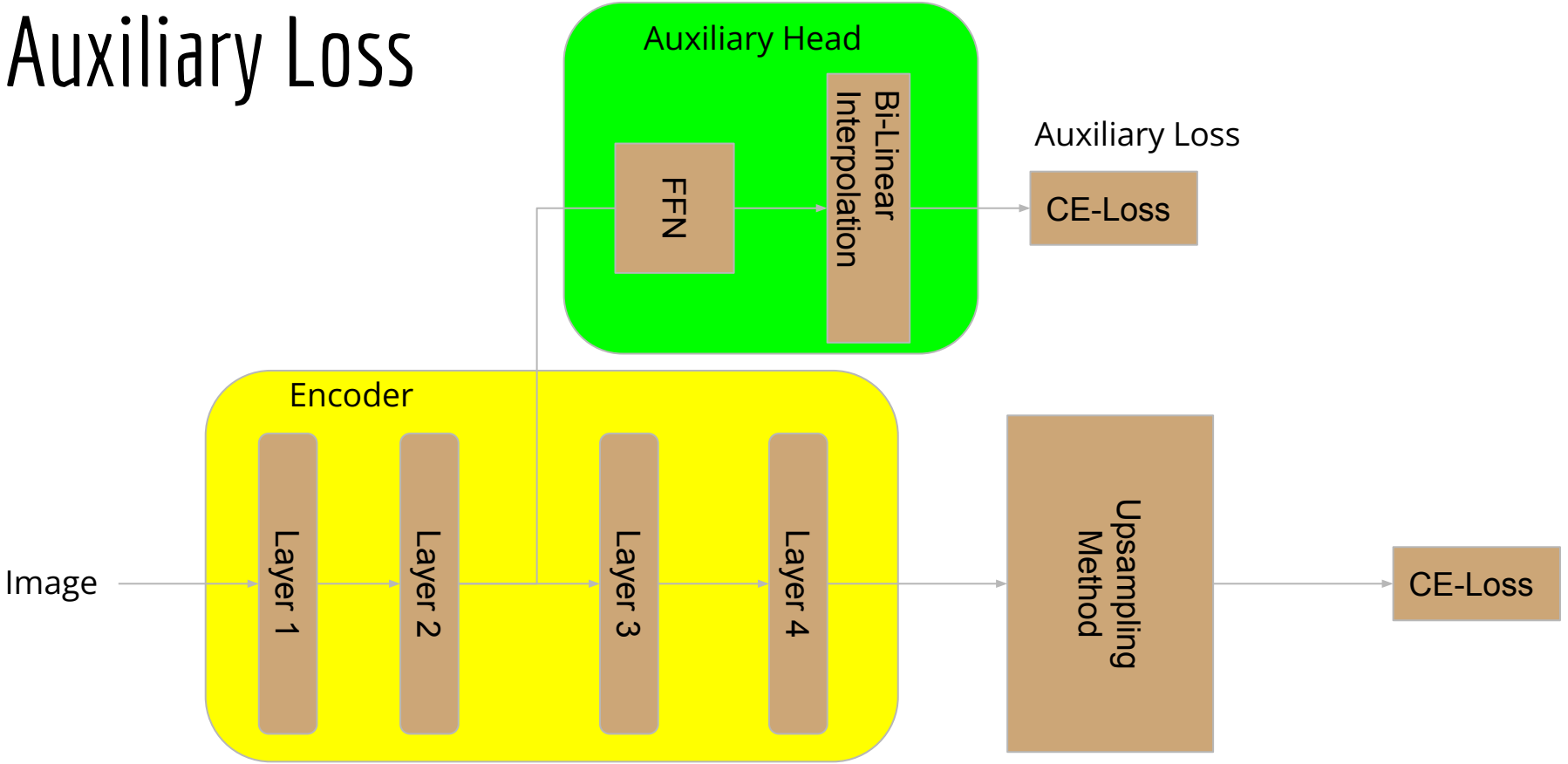


Loss Function

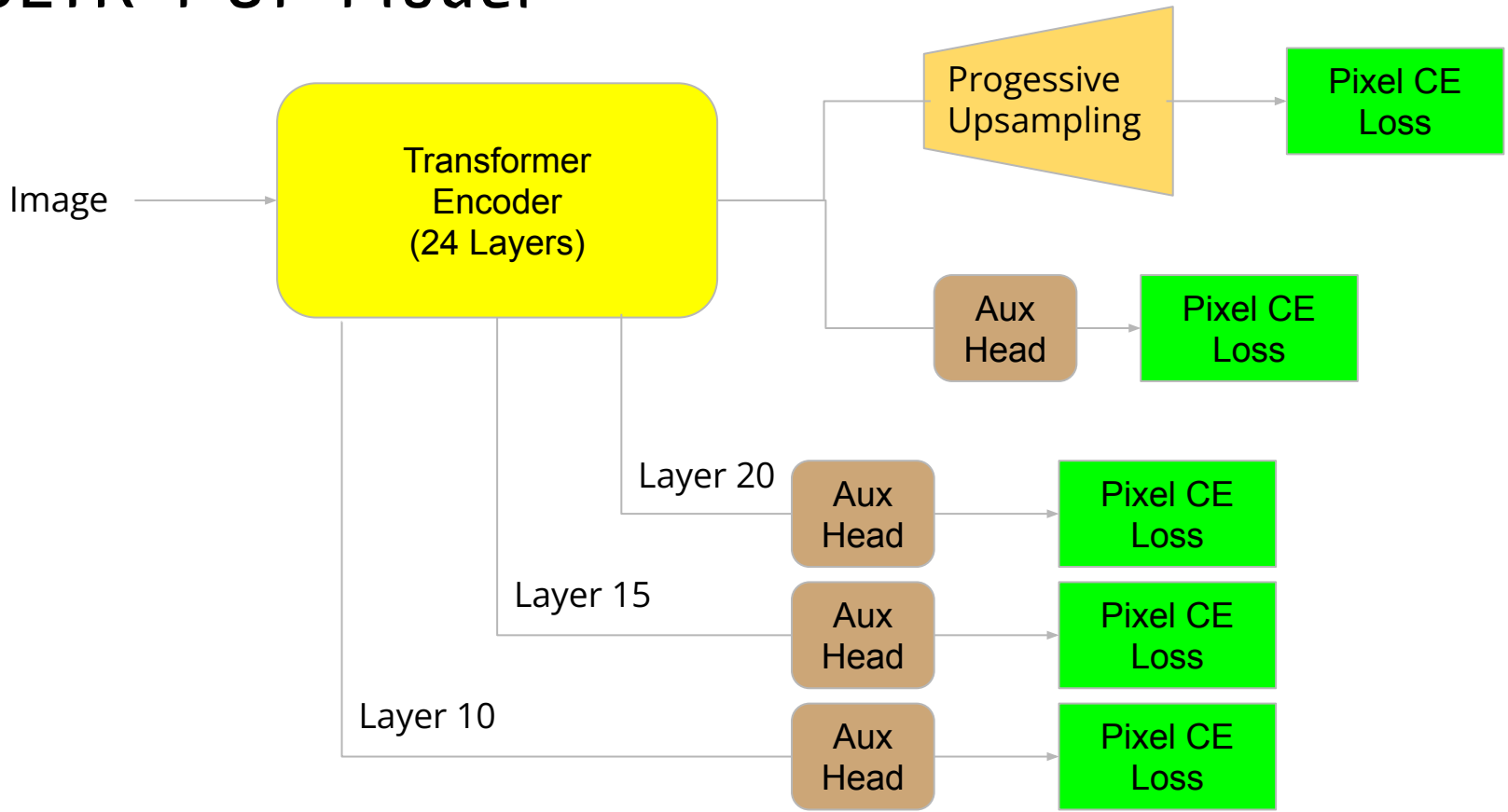
- Pixel-wise Cross-Entropy Loss
- Auxiliary Loss
  - Use outputs from different Encoder layers for loss
    - SETR-Naive: Layers 10,15,20
    - SETR-PUP: Layers 10,15,20,24
    - SETR-MLA: Layers 6,12,18,24
  - Outputs are sent through two-layer FFN before Bilinear Interpolation

$$\mathcal{L} = \mathcal{L}_{CE} + \gamma \sum_{i} \mathcal{L}_{Aux_i}$$

Auxiliary Loss



SETR-PUP Model



Experiment Datasets



Cityscapes

- 19 object classes, 3K train, 500 val, 1.5K test
- High resolution (2048x1024)



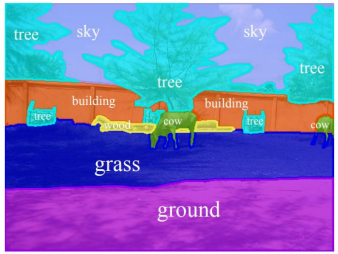


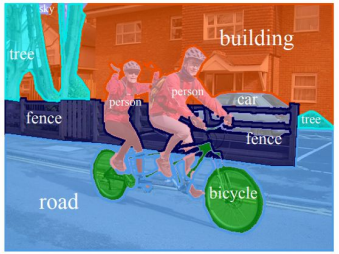
ADE20K

- 150 semantic classes, 20K train, 2K val, 3K test

PASCAL Context

- 60 classes, 5K train, 5K val





Experiment Details

- Augmentations
  - Random resize with ratio between 0.5 and 2
  - Random cropping (768, 512 and 480 for Cityscapes, ADE20K and Pascal Context)
  - Random horizontal flip
- Batch Size
  - 16 (ADE20K, Pascal)
  - 8 (Cityscapes)
- Polynomial Learning rate schedule
- SGD Optimizer
  - Momentum = 0.9
- Initial Learning Rate
  - 0.001 (ADE20K, Pascal)
  - 0.01 (Cityscapes)
- # Iterations defines length of training not epochs

Testing

- Single-Scale
  - Set scaling value
  - Crop -> Scale -> Horiz. Flip
- Multi-Scale
  - Multiple scales of an image can be used
  - Crop -> Pick scale from range/list -> Horiz. Flip
- Both utilize a sliding window to use all of the image
- Auxiliary Heads are ignored during testing

Comparison to state-of-the-art

Results on Cityscapes Validation

- New SOTA mean IOU (+0.65)

Method	Backbone	mIoU
FCN (40k, SS) [39]	ResNet-101	73.93
FCN (40k, MS) [39]	ResNet-101	75.14
FCN (80k, SS) [39]	ResNet-101	75.52
FCN (80k, MS) [39]	ResNet-101	76.61
PSPNet [60]	ResNet-101	78.50
DeepLab-v3 [10] (MS)	ResNet-101	79.30
NonLocal [48]	ResNet-101	79.10
CCNet [25]	ResNet-101	80.20
GCNet [4]	ResNet-101	78.10
Axial-DeepLab-XL [47] (MS)	Axial-ResNet-XL	81.10
Axial-DeepLab-L [47] (MS)	Axial-ResNet-L	81.50
SETR-*PUP* (40k, SS)	T-Large	78.39
SETR-*PUP* (40k, MS)	T-Large	81.57
SETR-*PUP* (80k, SS)	T-Large	79.34
SETR-*PUP* (80k, MS)	T-Large	**82.15**

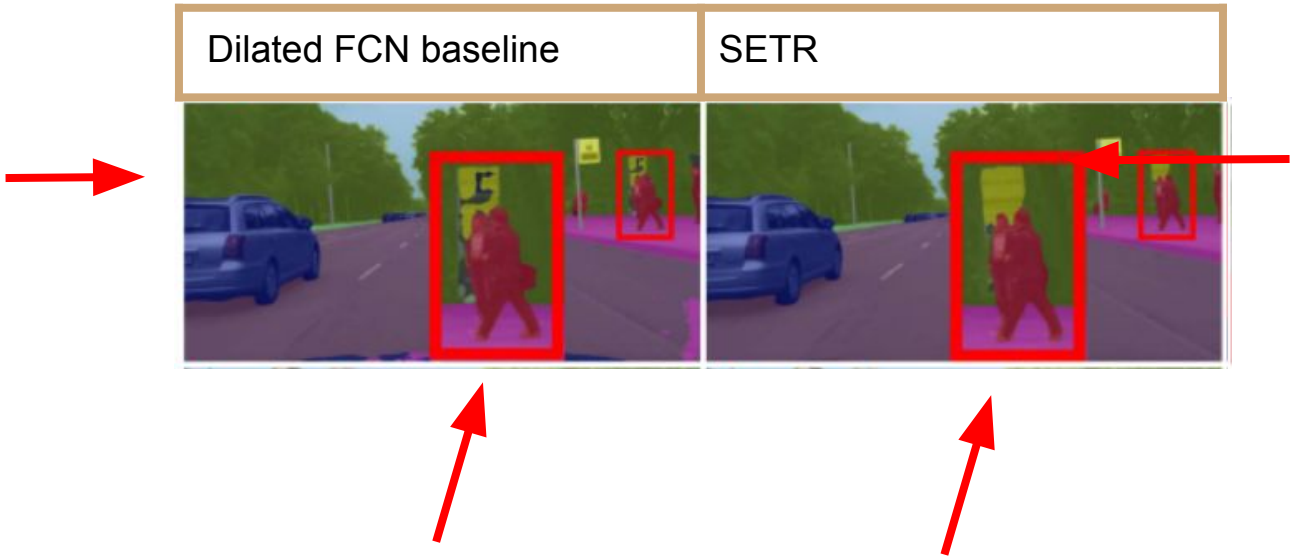
Comparison to state-of-the-art

Results on Cityscape Test

- (-0.26) behind SOTA
- Best result uses fine and coarse annotated data

Method	Backbone	mIoU
PSPNet [60]	ResNet-101	78.40
DenseASPP [50]	DenseNet-161	80.60
BiSeNet [52]	ResNet-101	78.90
PSANet [61]	ResNet-101	80.10
DANet [18]	ResNet-101	81.50
OCNet [55]	ResNet-101	80.10
CCNet [25]	ResNet-101	81.90
Axial-DeepLab-L [47]	Axial-ResNet-L	79.50
Axial-DeepLab-XL [47]	Axial-ResNet-XL	79.90
SETR-*PUP* (100k)	T-Large	81.08
SETR-*PUP*‡	T-Large	81.64

Qualitative Results on Cityscapes



Comparison to state-of-the-art

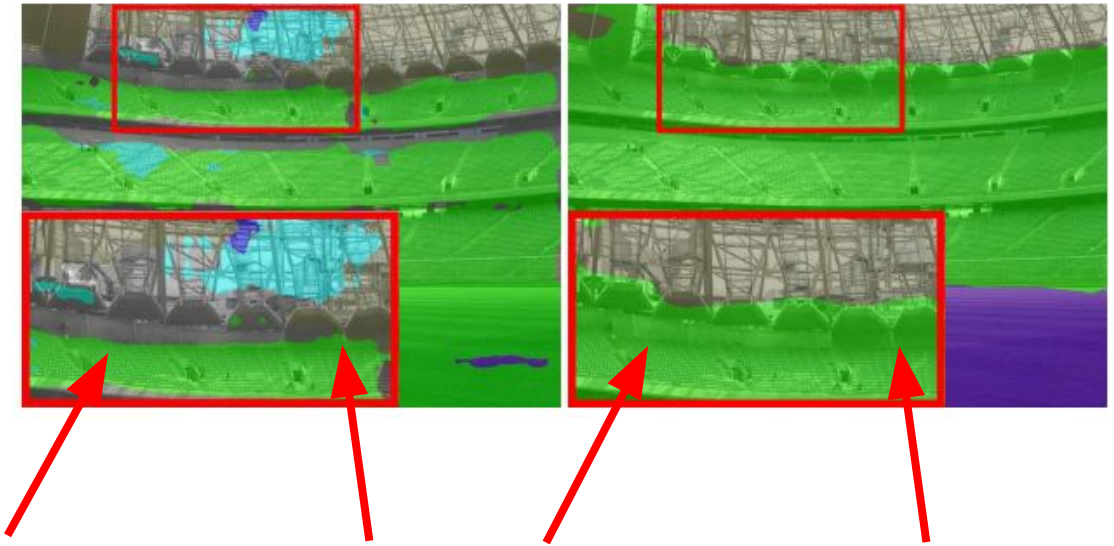
Results on ADE20K

- Almost every SETR model beats SOTA (+3.88)

Method	Pre	Backbone	#Params	mIoU
FCN (160k, SS) [39]	1K	ResNet-101	68.59M	39.91
FCN (160k, MS) [39]	1K	ResNet-101	68.59M	41.40
CCNet [25]	1K	ResNet-101	-	45.22
Strip pooling [23]	1K	ResNet-101	-	45.60
DANet [18]	1K	ResNet-101	69.0M	45.30
OCRNet [54]	1K	ResNet-101	71.0M	45.70
UperNet [49]	1K	ResNet-101	86.0M	44.90
Deeplab V3+ [11]	1K	ResNet-101	63.0M	46.40
SETR-*Naïve* (160k, SS)	21K	T-Large	305.67M	48.06
SETR-*Naïve* (160k, MS)	21K	T-Large	305.67M	48.80
SETR-*PUP* (160k, SS)	21K	T-Large	318.31M	48.58
SETR-*PUP* (160k, MS)	21K	T-Large	318.31M	50.09
SETR-*MLA* (160k, SS)	21K	T-Large	310.57M	48.64
SETR-*MLA* (160k, MS)	21K	T-Large	310.57M	**50.28**
SETR-*PUP-DeiT* (160k, SS)	1K	T-Base	97.64M	46.34
SETR-*PUP-DeiT* (160k, MS)	1K	T-Base	97.64M	47.30
SETR-*MLA-DeiT* (160k, SS)	1K	T-Base	92.59M	46.15
SETR-*MLA-DeiT* (160k, MS)	1K	T-Base	92.59M	47.71

Qualitative Results on ADE20K

Dilated FCN baseline	SETR



Comparison to state-of-the-art

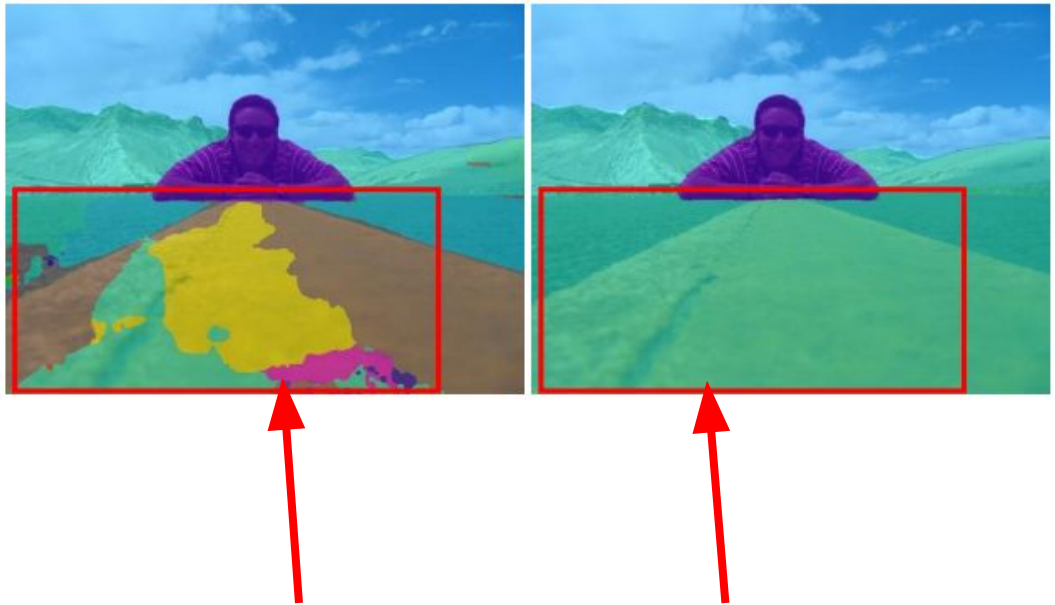
Pascal Context

- Improvement on baseline +6.97 to +10.09
- New SOTA on mean IOU (+1.13)

Method	Pre	Backbone	mIoU
FCN (80k, SS) [39]	1K	ResNet-101	44.47
FCN (80k, MS) [39]	1K	ResNet-101	45.74
DANet [18]	1K	ResNet-101	52.60
EMANet [31]	1K	ResNet-101	53.10
SVCNet [16]	1K	ResNet-101	53.20
Strip pooling [23]	1K	ResNet-101	54.50
GFFNet [30]	1K	ResNet-101	54.20
APCNet [19]	1K	ResNet-101	54.70
SETR-*Naïve* (80k, SS)	21K	T-Large	52.89
SETR-*Naïve* (80k, MS)	21K	T-Large	53.61
SETR-*PUP* (80k, SS)	21K	T-Large	54.40
SETR-*PUP* (80k, MS)	21K	T-Large	55.27
SETR-*MLA* (80k, SS)	21K	T-Large	54.87
SETR-*MLA* (80k, MS)	21K	T-Large	**55.83**
SETR-*PUP-DeiT* (80k, SS)	1K	T-Base	52.71
SETR-*PUP-DeiT* (80k, MS)	1K	T-Base	53.71
SETR-*MLA-DeiT* (80k, SS)	1K	T-Base	52.91
SETR-*MLA-DeiT* (80k, MS)	1K	T-Base	53.74

Qualitative Results on Pascal Context

Dilated FCN baseline	SETR



Qualitative Results: Attention Map

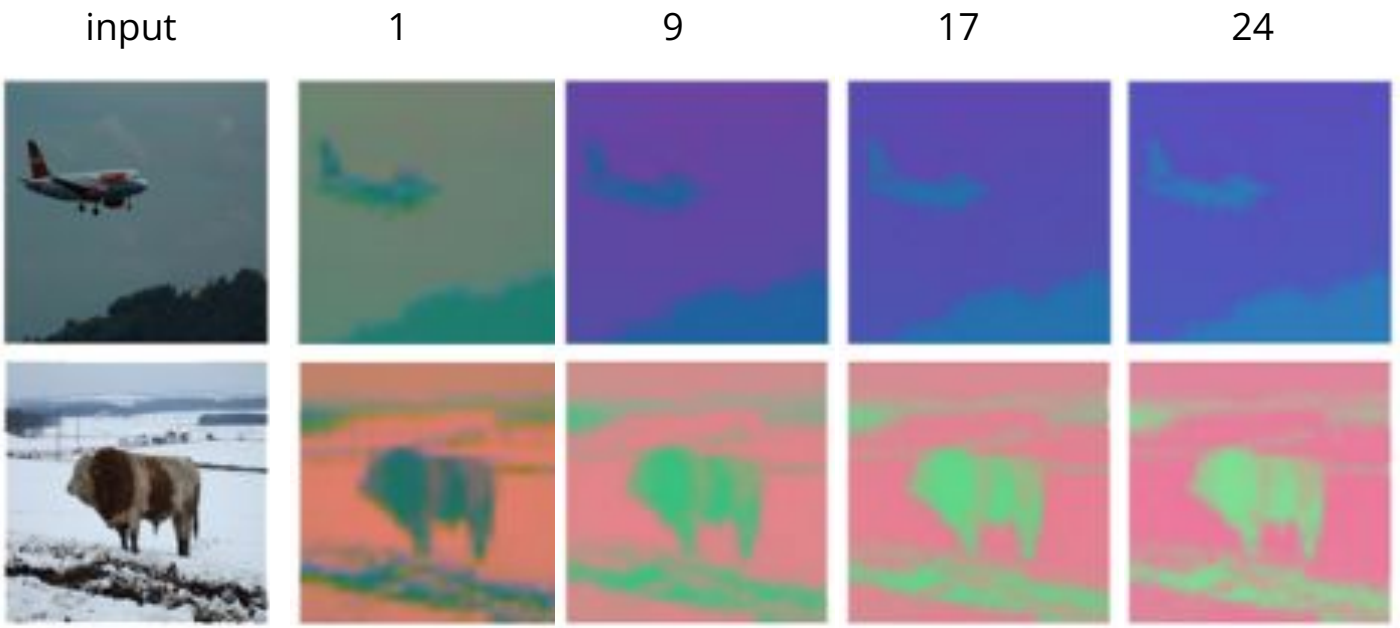
SETR trained on Pascal Context



Qualitative Results: Features

Output feature Visualizations of SETR on Pascal Context

Layers:



Ablation: SETR Variants

Cityscape benchmark

Method	Pre	Backbone	#Params	40k	80k
FCN [39]	1K	R-101	68.59M	73.93	75.52
Semantic FPN [39]	1K	R-101	47.51M	-	75.80
Hybrid-Base	R	T-Base	112.59M	74.48	77.36
Hybrid-Base	21K	T-Base	112.59M	76.76	76.57
Hybrid-DeiT	21K	T-Base	112.59M	77.42	78.28
SETR-*Naïve*	21K	T-Large	305.67M	77.37	77.90
SETR-*MLA*	21K	T-Large	310.57M	76.65	77.24
SETR-*PUP*	21K	T-Large	318.31M	78.39	79.34
SETR-*PUP*	R	T-Large	318.31M	42.27	-
SETR-*Naïve-Base*	21K	T-Base	87.69M	75.54	76.25
SETR-*MLA-Base*	21K	T-Base	92.59M	75.60	76.87
SETR-*PUP-Base*	21K	T-Base	97.64M	76.71	78.02
SETR-*Naïve-DeiT*	1K	T-Base	87.69M	77.85	78.66
SETR-*MLA-DeiT*	1K	T-Base	92.59M	78.04	78.98
SETR-*PUP-DeiT*	1K	T-Base	97.64M	**78.79**	**79.45**

Ablation: SETR Variants

ADE20K benchmark

Method	Pre	Backbone	#Params	mIoU
SETR-*Naïve* (160k, SS)	21K	T-Large	305.67M	48.06
SETR-*Naïve* (160k, MS)	21K	T-Large	305.67M	48.80
SETR-*PUP* (160k, SS)	21K	T-Large	318.31M	48.58
SETR-*PUP* (160k, MS)	21K	T-Large	318.31M	50.09
SETR-*MLA* (160k, SS)	21K	T-Large	310.57M	48.64
SETR-*MLA* (160k, MS)	21K	T-Large	310.57M	**50.28**
SETR-*PUP-DeiT* (160k, SS)	1K	T-Base	97.64M	46.34
SETR-*PUP-DeiT* (160k, MS)	1K	T-Base	97.64M	47.30
SETR-*MLA-DeiT* (160k, SS)	1K	T-Base	92.59M	46.15
SETR-*MLA-DeiT* (160k, MS)	1K	T-Base	92.59M	47.71

Ablation: SETR Variants FCN Comparison

Comparison to FCN with different pre-training on ADE20K and Cityscapes

Method	Pre	Backbone	ADE20K	Cityscapes
FCN [39]	1K	R-101	39.91	73.93
FCN	21K	R-101	42.17	76.38
SETR-*MLA*	21K	T-Large	**48.64**	76.65
SETR-*PUP*	21K	T-Large	48.58	78.39
SETR-*MLA-DeiT*	1K	T-Large	46.15	78.98
SETR-*PUP-DeiT*	1K	T-Large	46.24	**79.45**

Conclusion

- Alternative method for semantic segmentation (sequence to sequence prediction network)
- Eliminate the reliance on FCN
- Shown Transformers can model global context at every stage of feature learning
- State of the art on ADE20K and Pascal Context
- Competitive results on Cityscapes

References

1. Zheng, Sixiao, et al. "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
2. *Long, J., Shelhamer, E., Darrel, T., & UC Berkeley. (n.d.). Fully Convolutional Networks for Semantic Segmentation. CVPR 2021 open access repository. Retrieved April 15, 2022, from https://openaccess.thecvf.com/CVPR2021*
3. *Dosovitskiy, Alexey, et al. (n.d.). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. " ICLR 2021 Conference Paper1909 Comment*
4. Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand- alone axial-attention for panoptic segmentation. In *ECCV*, 2020.
5. *Visual object classes challenge 2012 (VOC2012)*. The PASCAL Visual Object Classes Challenge 2012 (VOC2012). (n.d.). Retrieved April 18, 2022, from http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit