Panoptic FCN

October 6th, 2021

Overview

Model

Experiments

Recall:

Panoptic Segmentation

- Study of ***stuff*** <u>and</u> ***things***
- Assign one class label and instance id to each pixel in an image
- Evaluated by Panoptic Quality (PQ)

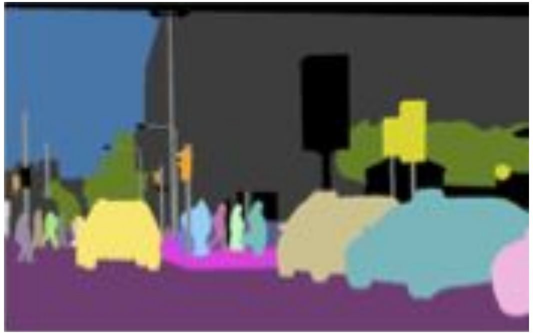


Figure from: Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9404-9413).

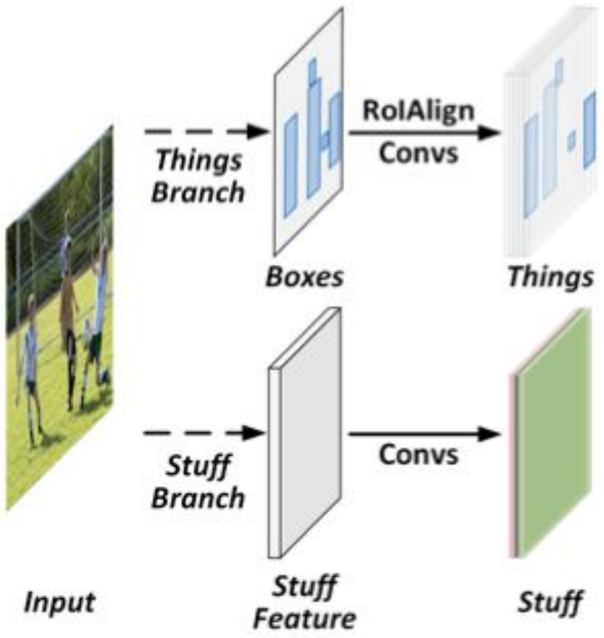
Difficulty of unifying segmentation

- Countable things are discovered through instance-aware features to distinguish entities
- Stuff regions are found through semantically consistent features

Figure from: https://www.scientificamerican.com/article/should-kids-be-allowed-to-play-soccer/

Separate branches

- Differing feature needs led to models with separate branches
- Things were addressed by box-based and box-free branches
- Stuff was addressed by pixel-by-pixel branches



Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Separate branches

Panoptic FPN

- Mask R-CNN for things
- FCN for stuff

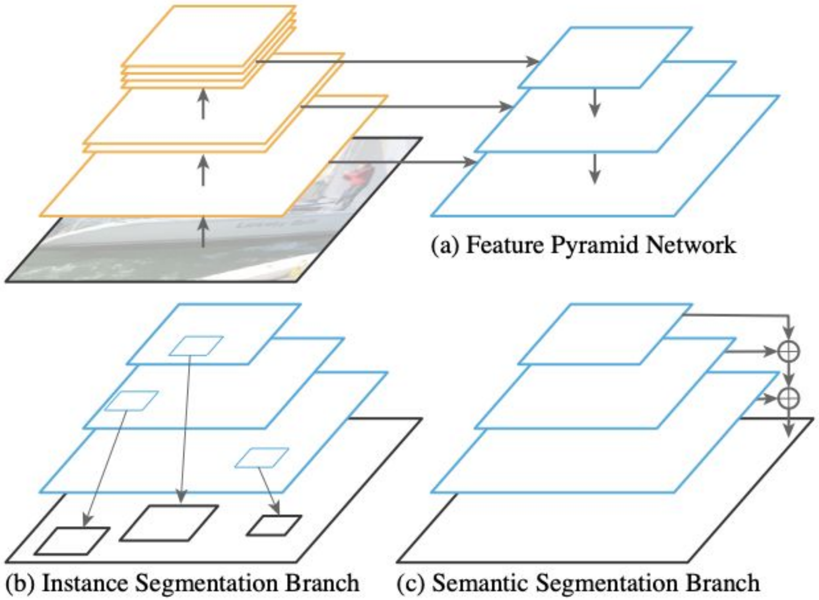


Figure from: Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

Separate branches

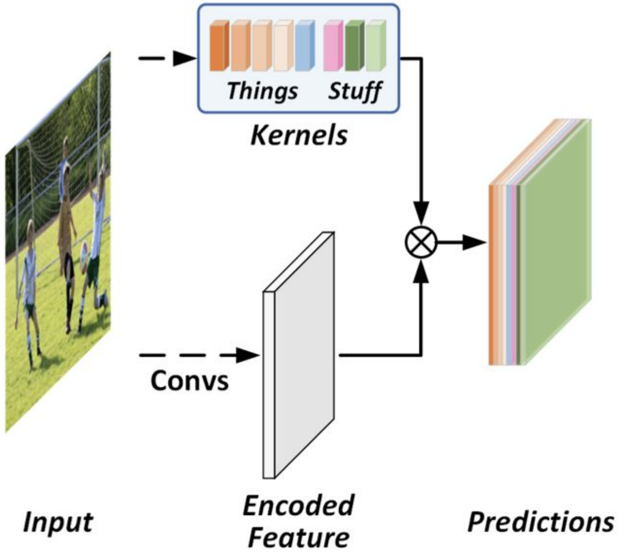
An ununified workflow

- Separate branches don’t handle prediction uniformly
- Not in the spirit of PS

Figure from: Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

Unification

- Represent things and stuff features in the same way
- Predict things and stuff together



Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Recall:

FCN for Semantic Segmentation

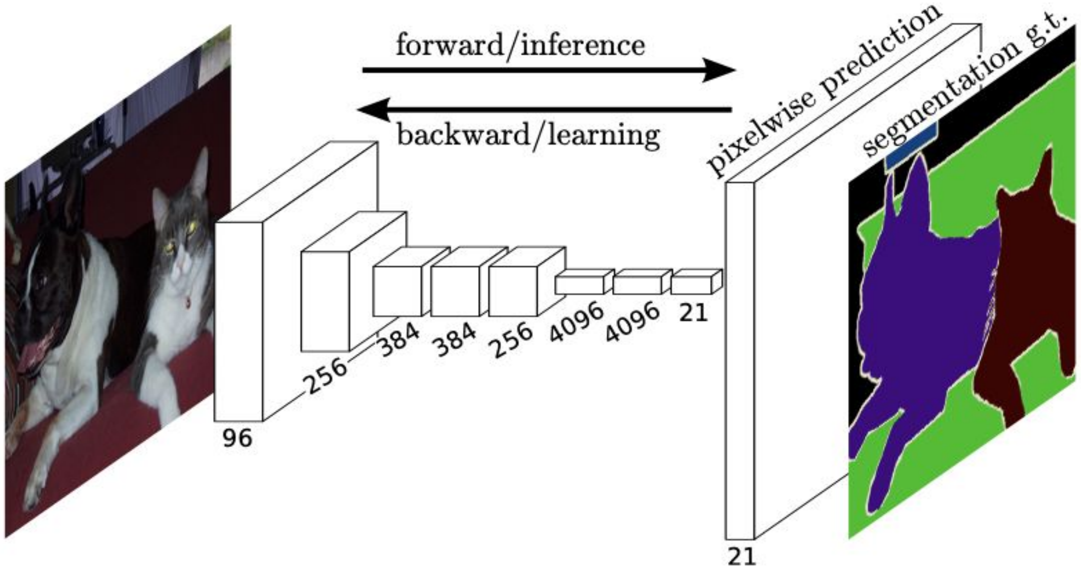


Figure from: Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Overview
Model
Experiments

Architecture

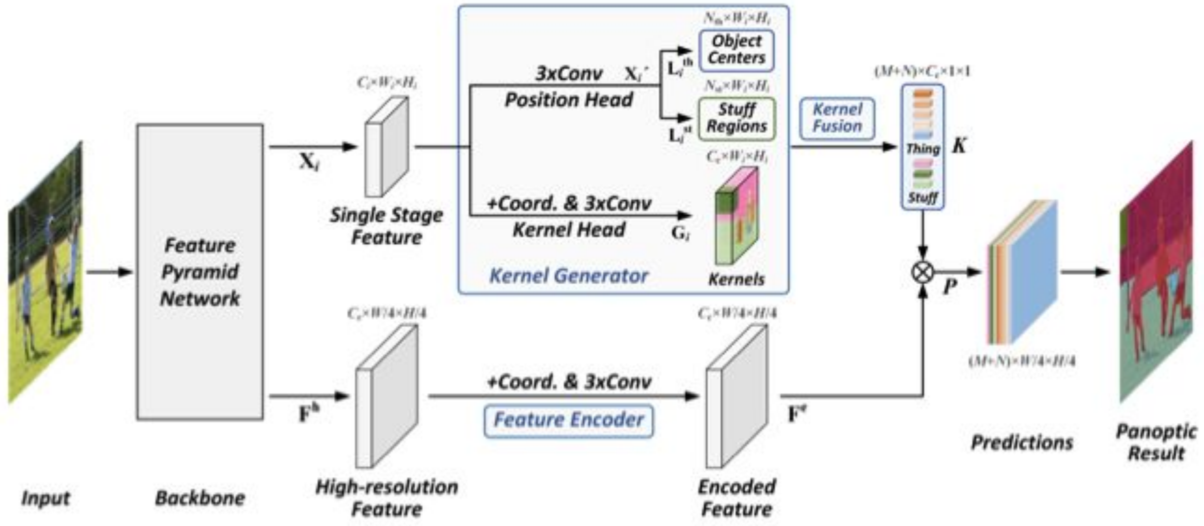


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture

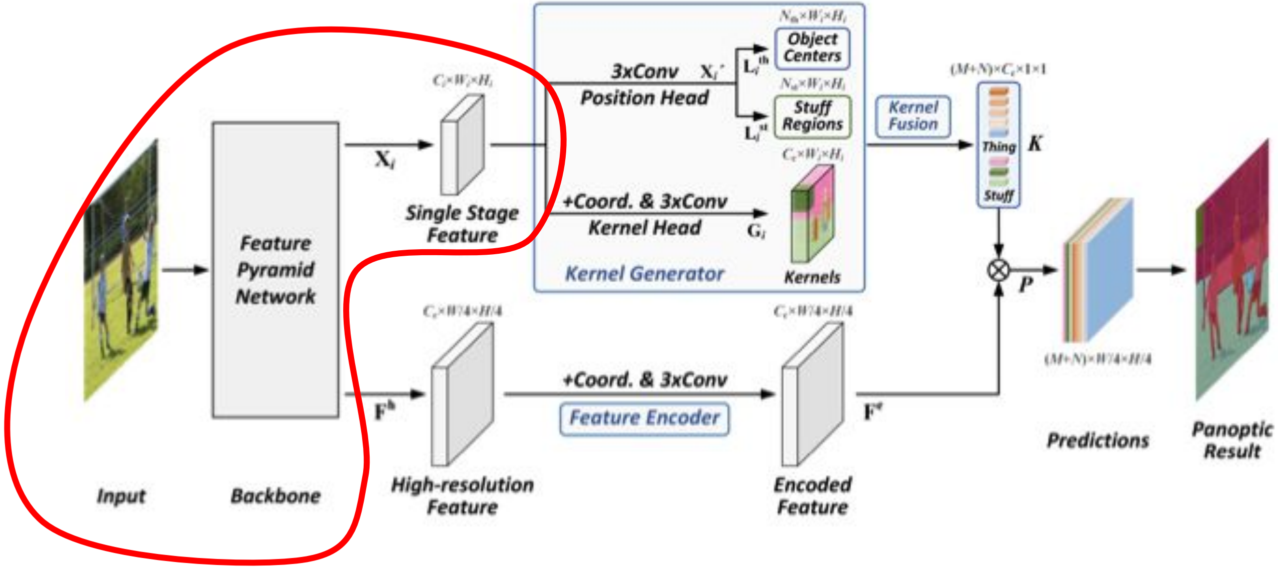


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture

Feature Pyramid Network

- FPNs proven to be a very effective feature extraction method
- Utilize FPN to help detect objects at different scales
- Compute feature map at each stage of the FPN

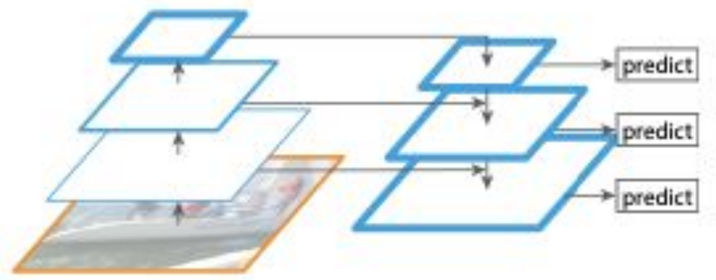


Figure from: Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

Architecture

Feature Pyramid Network

- Pass each feature map in separately to the Kernel Generator module

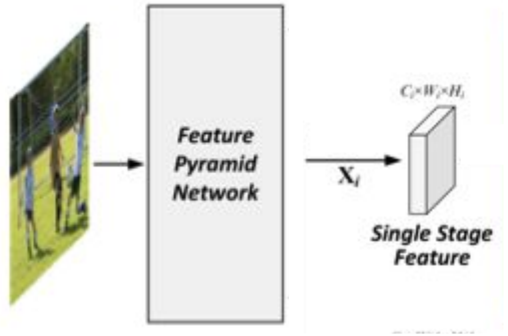


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture

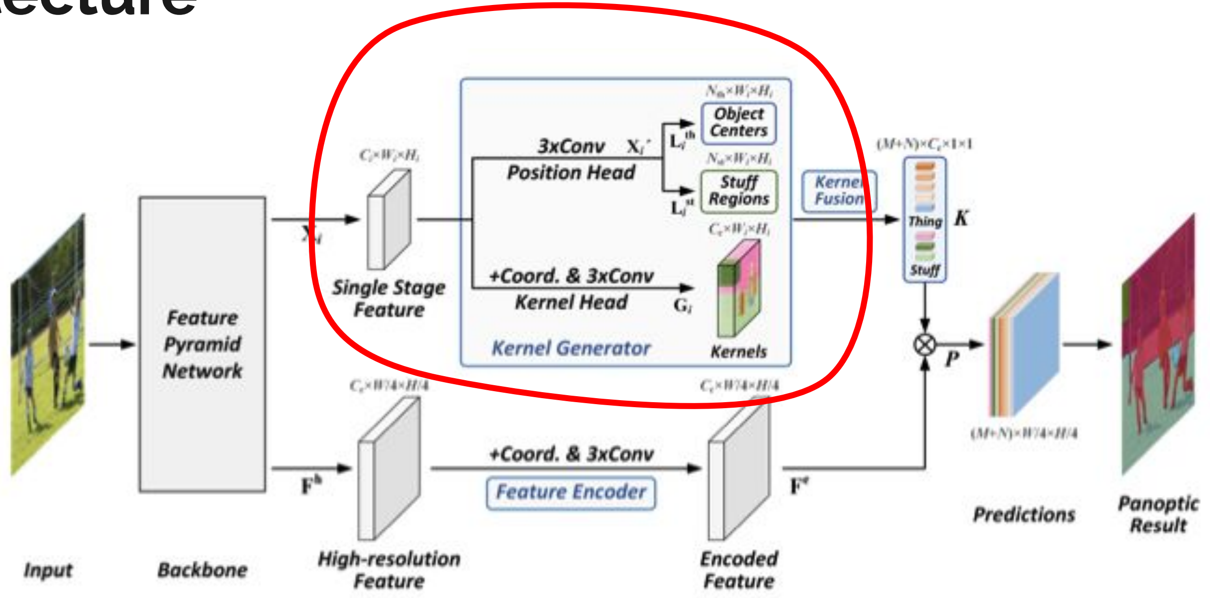


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture

Kernel Generator?

- Traditional convolutional layers uses static filters
- But this dynamically generates filters based on the current input
- See *Dynamic Filter Networks*

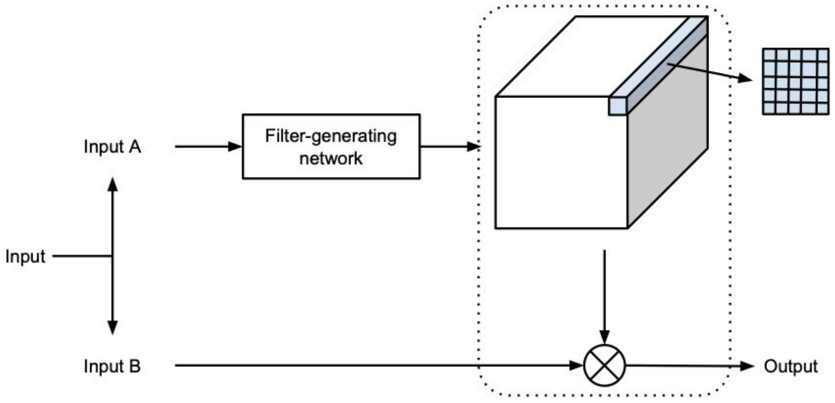


Figure from: Jia, X., De Brabandere, B., Tuytelaars, T., & Gool, L. V. (2016). Dynamic filter networks. *Advances in neural information processing systems*, *29*, 667-675.

Architecture

Why use dynamic filters?

- We can extract features specific to the objects in the image
- Can adjust number of output tensor channels for varying amounts of instances

Figures from:

1. https://coachart.org/blog/4-benefits-of-soccer-for-kids-with-adaptations-for-disability-inclusion/
2. https://www.activekids.com/sports/articles/the-most-inexpensive-sports-for-kids

Architecture

Kernel Generator

From each single stage feature, a...

- **Position head** performs localization and classification
- **Kernel head** generates kernel weights

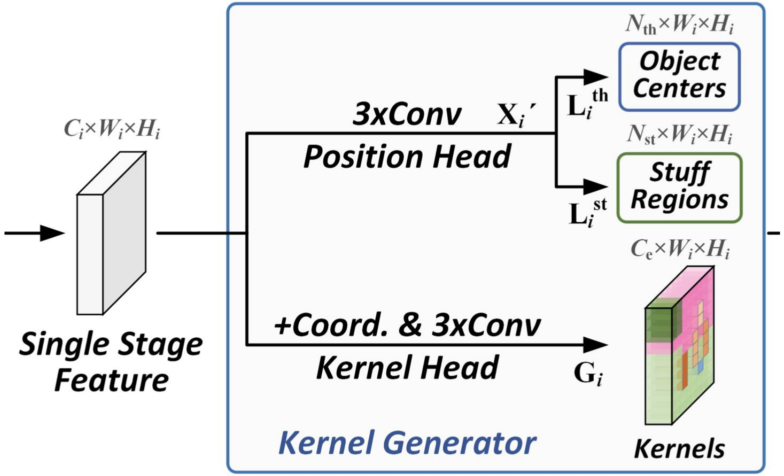


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture: Kernel Generator

Position Head

- Run input feature map through stacks of convolutions
- Generate a map for **object centers** and another for **stuff regions** through 2 branches

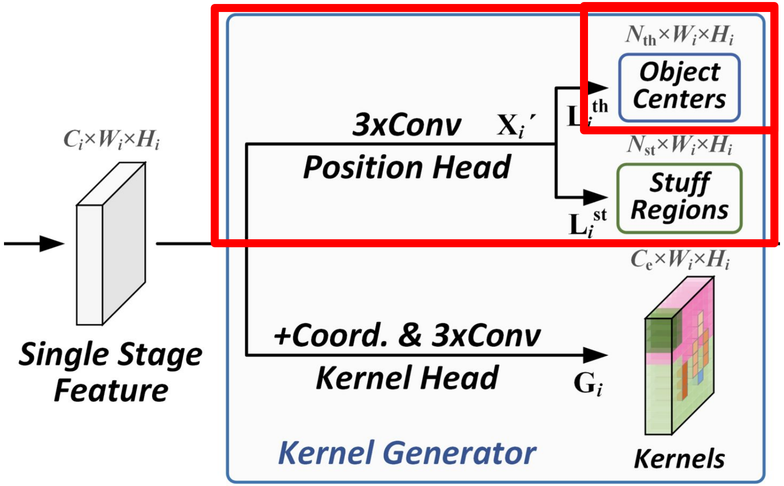


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Generator: Position Head

Object Centers

- Similar to *CenterNet*
  - Generates heat maps with the likelihood each pixel is an object center
  - Fully convolutional network
- Training requires us to generate ground truths

Figure from: Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Kernel Generator: Position Head

Object Centers: GTs

- Two approaches to get center keypoints from annotated images:
  - Center of mass for each mask
  - Center of bounding box

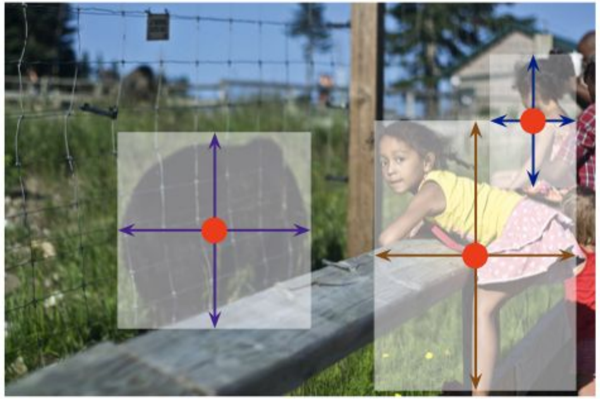


Figure from: Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Kernel Generator: Position Head

Object Centers: GTs

- Pass center keypoints to a Gaussian kernel to generate ground truth heat map

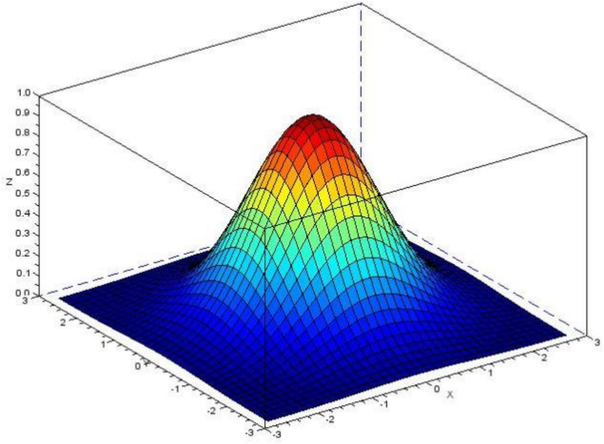


Figure from: https://zbigatron.com/generating-heatmaps-from-coordinates/

Kernel Generator: Position Head

Object Centers: Loss

$$\mathcal{L}_{\text{pos}}^{\text{th}} = \sum_i \underbrace{\text{FL}}_{\text{Focal Loss}}(\mathbf{L}_i^{\text{th}}, \mathbf{Y}_i^{\text{th}})/N_{\text{th}},$$

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Focal Loss

- Enhance *Cross-Entropy Loss* by reducing loss impact from well-classified examples
- Adds a tunable **focusing** parameter

Focusing parameter

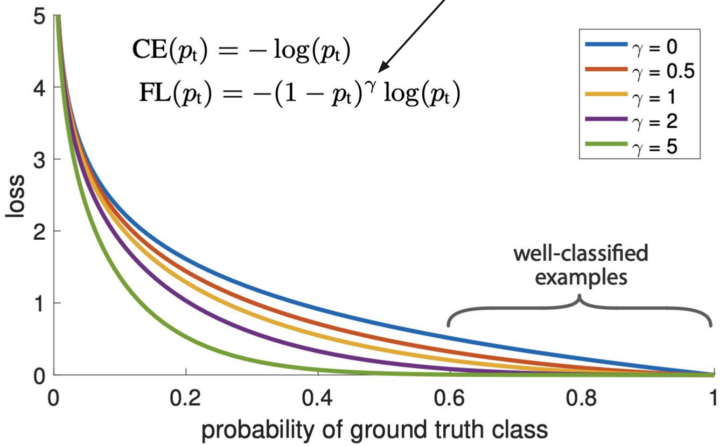


Figure from: Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

Kernel Generator: Position Head

Object Centers: Loss

Object Centers map

$$\mathcal{L}^{\mathrm{th}}_{\mathrm{pos}} = \sum_i \mathrm{FL}(\mathbf{L}^{\mathrm{th}}_i, \mathbf{Y}^{\mathrm{th}}_i)/N_{\mathrm{th}},$$

Focal Loss

Center keypoint heatmap

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture: Kernel Generator

Position Head

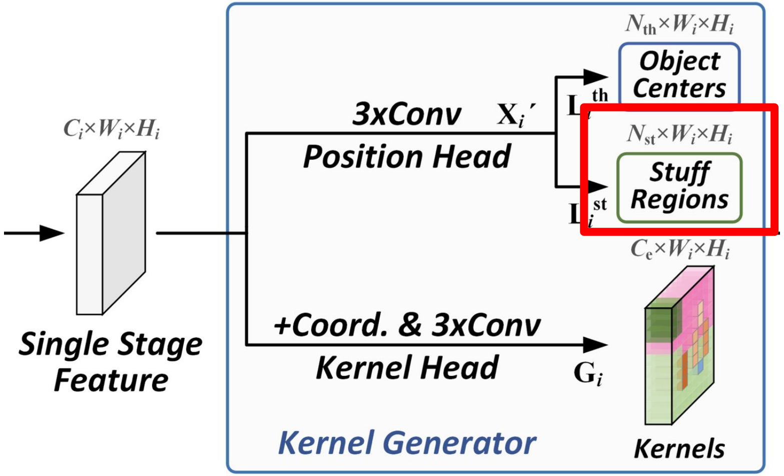


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Generator: Position Head

Stuff Regions

- Fully convolutional network
- Training requires us to generate ground truths
  - Bilinear interpolate semantic labels from the annotated images
  - Same resolution as feature map

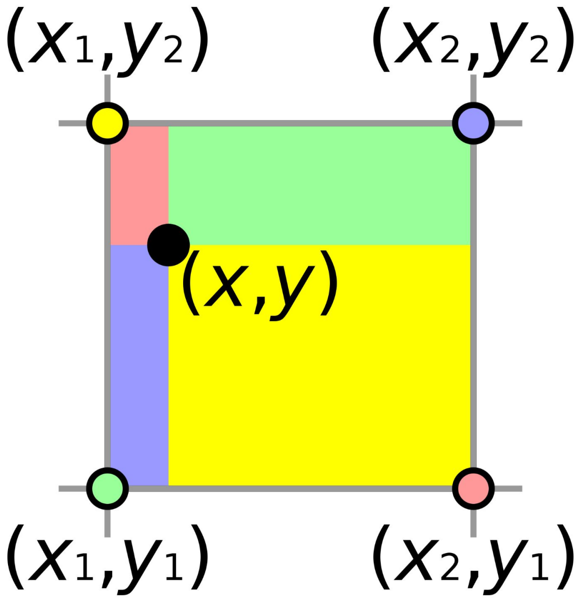


Figure from:
https://upload.wikimedia.org/wikipedia/commons/9/91/Bilinear_interpolation_visualisation.svg

Kernel Generator: Position Head

Stuff Regions: Loss

Stuff Regions map

$$\mathcal{L}_{\text{pos}}^{\text{st}} = \sum_i \text{FL}(\mathbf{L}_i^{\text{st}}, \mathbf{Y}_i^{\text{st}})/W_i H_i,$$

Focal Loss

Bilinear interpolated segmentation masks

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture: Kernel Generator

Position Head

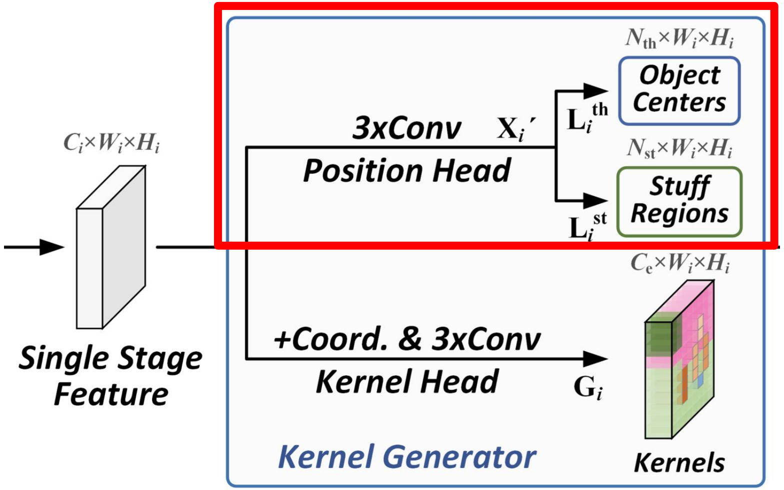


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Generator: Position Head

Multitask Loss

$$\mathcal{L}_{\mathrm{pos}} = \mathcal{L}_{\mathrm{pos}}^{\mathrm{th}} + \mathcal{L}_{\mathrm{pos}}^{\mathrm{st}}$$

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Generator: Position Head

Output

- Collect two sets of coordinates with corresponding labels for things and stuff
- Corresponding label will be the highest likelihood thing or stuff class for that point in the feature map
    - Must surpass threshold

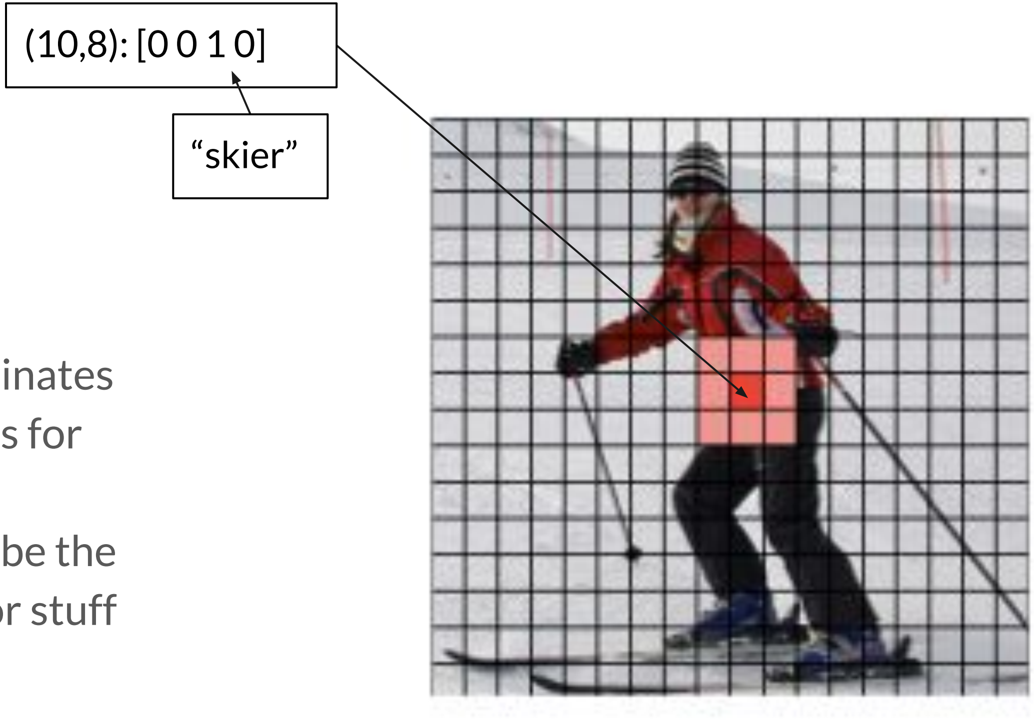


Figure from: Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Architecture: Kernel Generator

Kernel Head

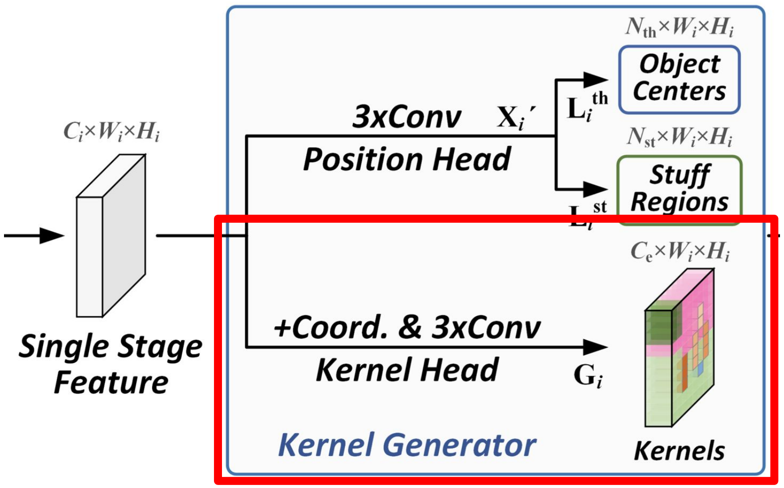


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Architecture: Kernel Generator

Kernel Head

- Concatenate coordinates of each feature
  - *CoordConv* showed this to improve results related to coordinates in ConvNets
- Run map with coordinates through stacks of convolutions

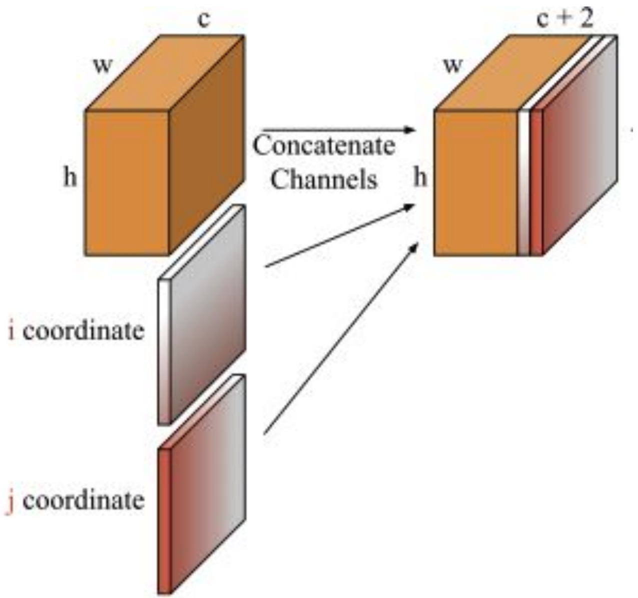
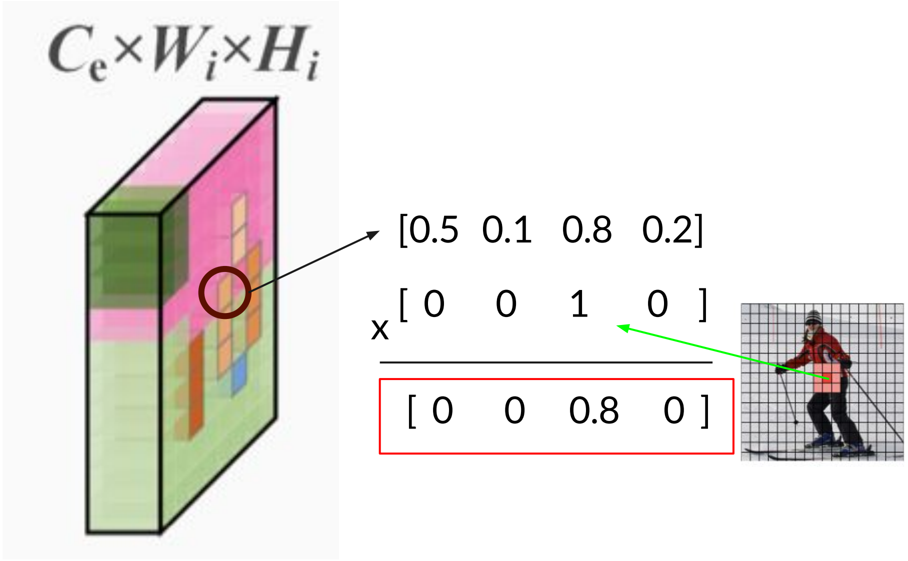


Figure from: Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., & Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*.

Architecture: Kernel Generator

Kernel Head

- Select weights from the feature map we just generated
- Find matching coordinates in the two sets created for things and stuff
- Create two separate kernel weight maps for things and stuff

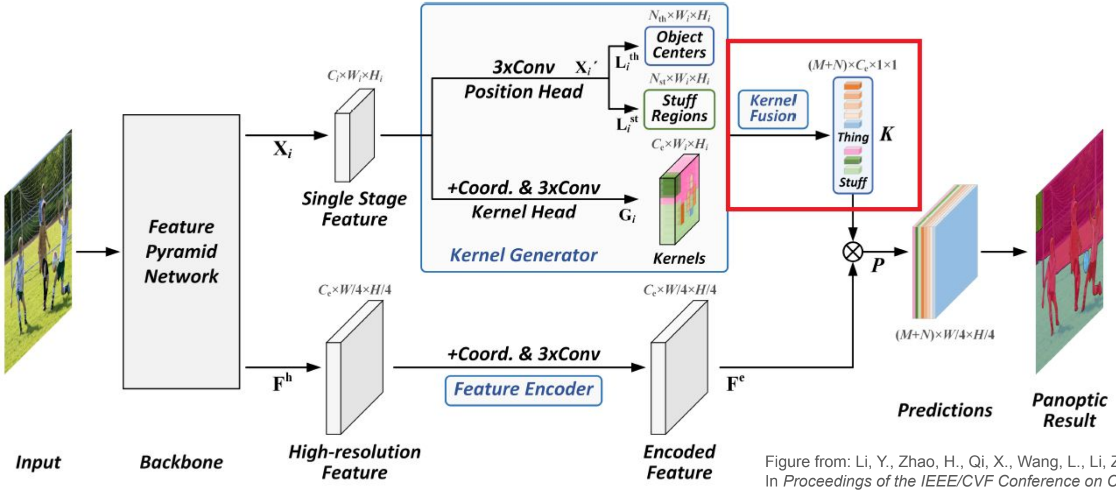


Figures from:

- Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.
- Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Fusion

To ensure instance awareness and semantic-consistency for things and stuff, respectively.



Merges repetitive kernel weights from multiple stages before final instance generation

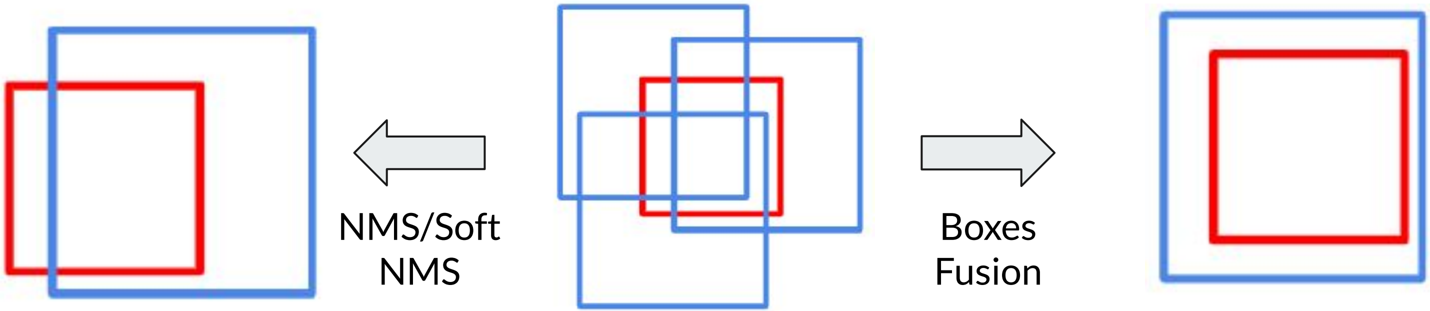
How?

Weighted Boxes Fusion like procedure

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Weighted Boxes Fusion

Note: Not used in Panoptic FCN



BB from multiple stages

- Filters out BB with low IoU
- Eliminates boxes with low confidence scores

https://arxiv.org/abs/1910.13302

- Filters out BB with low IoU
- Weighted Average of filtered BB proposals

Figure from: Weighted boxes fusion: Ensembling boxes from different object detection models Roman Solovyev, Weimin Wang, Tatiana Gabruseva

Kernel Fusion

How BB are in weighted box fusion, Kernel Weights are in Kernel Fusion

$$G = \{G_1, G_2, \ldots, G_j, \ldots G_{m+n}\} \longrightarrow G^{th} = \{G^{th}_1, G^{th}_2, \ldots, G^{th}_j, \ldots G^{th}_m\} + G^{st} = \{G^{st}_1, G^{st}_2, \ldots, G^{st}_j, \ldots G^{st}_n\}$$

All Kernel Weights

Kernel Weights for things

Kernel Weights for stuff

Kernel Weights when convolved with high resolution features gives predictions & segmentation outputs

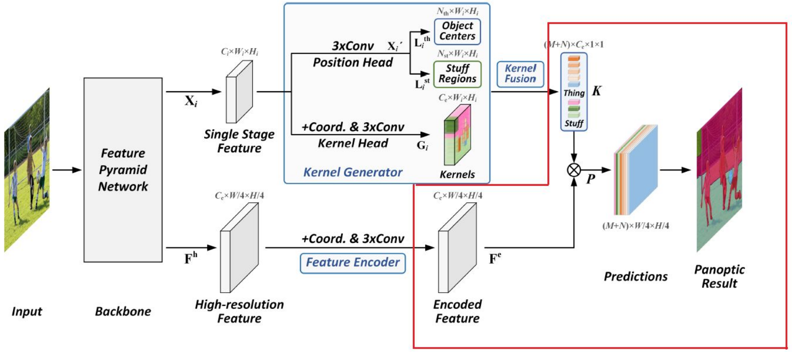


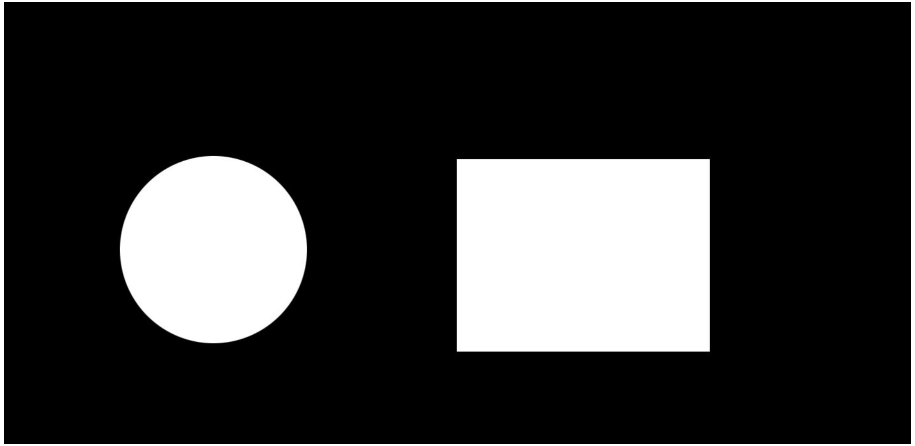
Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Fusion

Intuition and illustration

$$G = \{G_1, G_2, \ldots, G_j, \ldots G_{M+N}\} \rightarrow G^{th} = \{G^{th}_1, G^{th}_2, \ldots, G^{th}_j, \ldots G^{th}_M\} + G^{st} = \{G^{st}_1, G^{st}_2, \ldots, G^{st}_j, \ldots G^{st}_N\}$$

Original Image



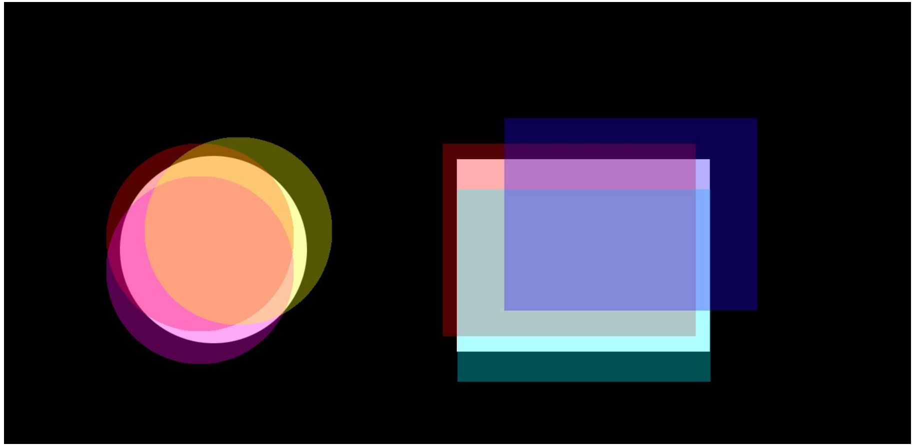
Let’s visualize how these kernel weights looks like after convolution

Kernel Fusion

Intuition

$$G^{th} = \{G^{th}{}_{1}, G^{th}{}_{2}, \ldots, G^{th}{}_{j}, \ldots G^{th}{}_{M}\}$$

Predicted Kernels (Visualized when convolved)



Kernel Fusion

Fusion Steps

$$G^{th} = \{G^{th}{}_{1}, G^{th}{}_{2}, \ldots, G^{th}{}_{j}, \ldots G^{th}{}_{M}\}$$

Step1: Create 2 Empty sets

G’ = {Set of clusters}

K = {Set of fused kernel weights}

Step 2: Iterate through the set G and update G’

Kernel Fusion

Step 2: Iterate through the set G and update G'

G' = {Set of clusters}

Top scoring kernel weight

How to identify a cluster?

$$G'_j = \{G_m : \mathrm{ID}(G_m) = \mathrm{ID}(G_j)\}$$

How is ID determined?

Things	**Stuff**
If the cosine similarity surpasses a given threshold	All kernel weights which share the same category are marked as one ID

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Fusion

Step 2: Iterate through the set G and update G'

$$G'_j = \{G_m : \mathrm{ID}(G_m) = \mathrm{ID}(G_j)\}$$

$G^{th} = \{G^{th}_{1}, G^{th}_{2}, \ldots, \ldots G^{th}_{8}\}$

G' = {Set of clusters}

G'_{th} = {{G1, G6, G8}, {G3, G4, G5}}

Assume Thres = 0.9

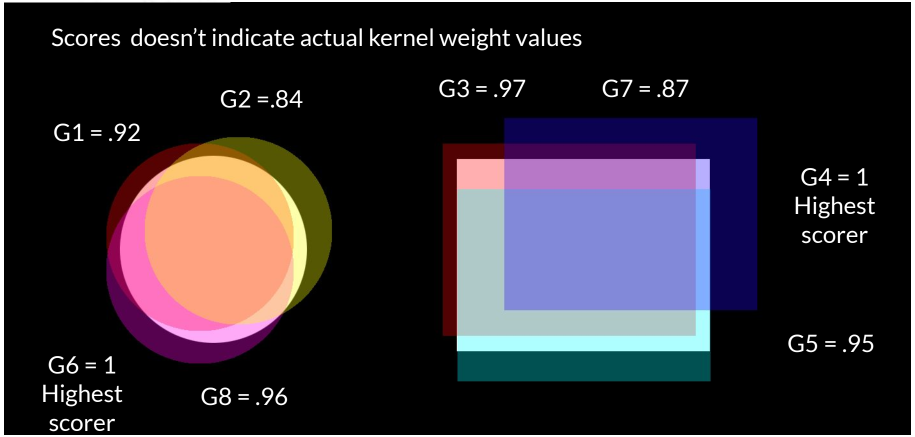


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Fusion

Step 3: Generate final Kernel weights

$G'_{th} = \{\{G1, G6, G8\}, \{G3, G4, G5\}\}$

$$K_j = \mathrm{AvgCluster}(G'_j),$$

$K^{th} = \{K1, K2\}$

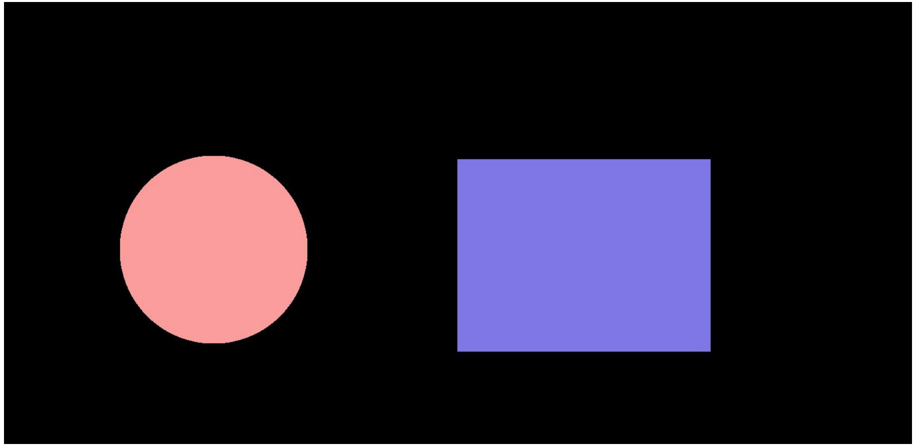
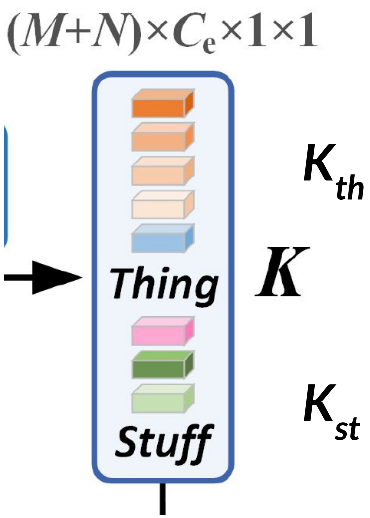


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Kernel Fusion



$$K^{st} = \{K^{st}_1, K^{st}_2, \dots, K^{st}_j, \dots K^{st}_N\}$$

$$K^{th} = \{K^{th}_1, K^{th}_2, \dots, K^{th}_j, \dots K^{th}_M\}$$

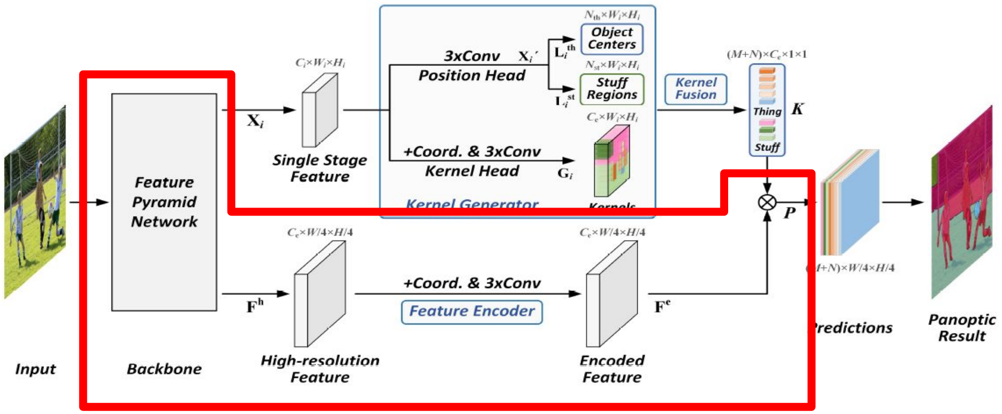
$$K = \{K_1, K_2, \dots, K_j, \dots K_{M+N}\}$$

Each kernel weight can be viewed as an embedding of a single object or stuff

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Overview

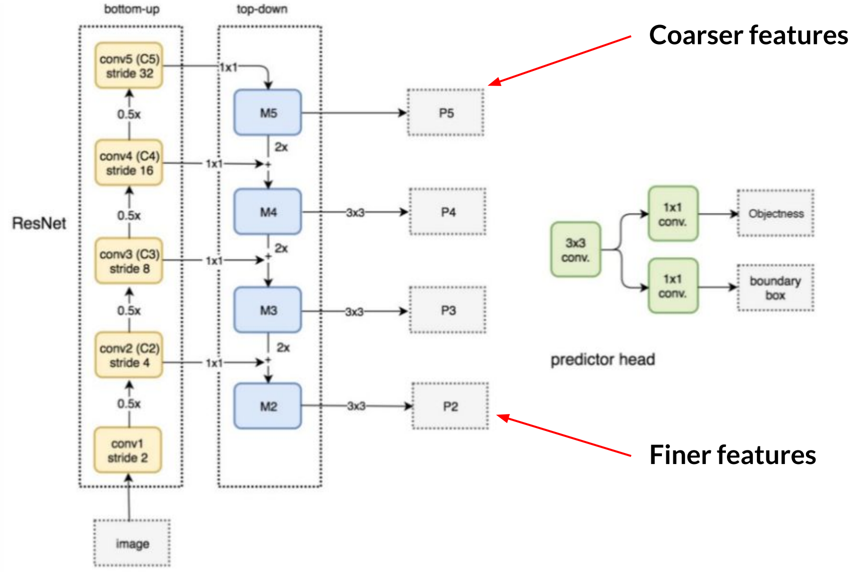


1. **Which output from the FPN network to use for high resolution feature extraction?**
2. Why encode position information?
3. The convolution step

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Which output from the FPN network to use for high resolution feature extraction?



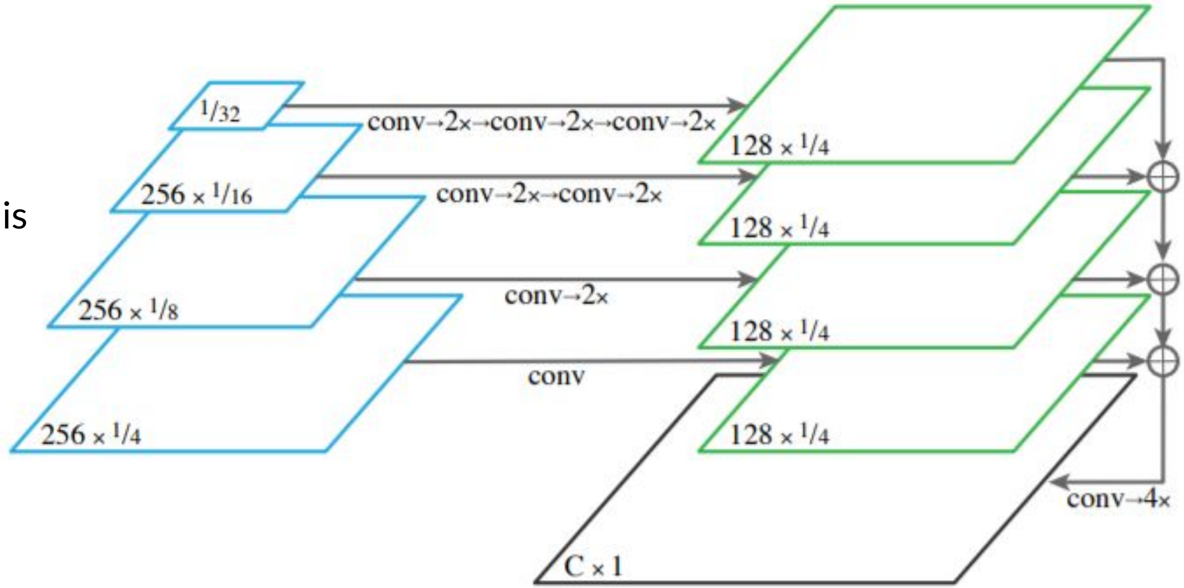
1. P2 stage feature
2. Summed up feature from all stages
3. **Features from semantic FPN ??**

Figure_Source:https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c

Feature Encoder

Semantic FPN

- Each stage of downsampling is upsampled to ¼ size
- Outputs from feature pyramid is element wise summed up



Figure_Source:Figure_Source:https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c

Feature Encoder

Which output from the FPN network to use for high resolution feature extraction?

1. P2 stage feature
2. Summed up feature from all stages
3. Features from semantic FPN

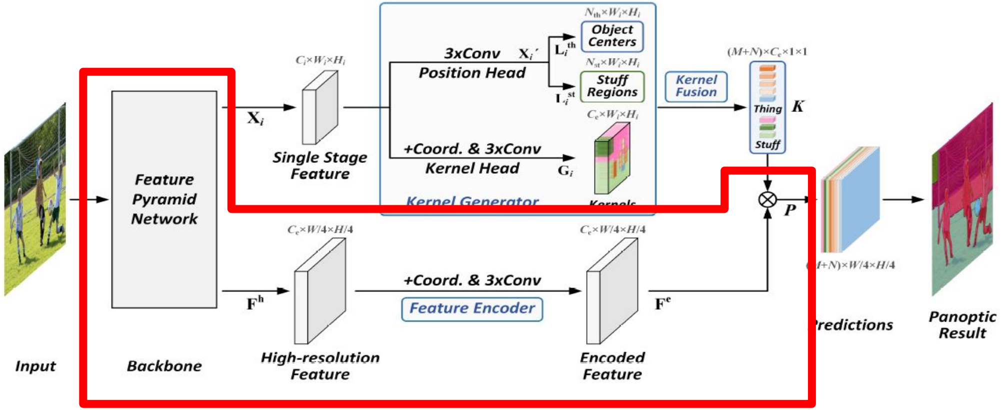
feature type	PQ	PQ^{th}	PQ^{st}	AP	mIoU
FPN-P2	40.6	46.0	32.4	31.6	41.3
FPN-Summed	40.5	46.0	32.1	31.7	41.1
Semantic FPN [18]	**41.3**	**46.9**	**32.9**	**32.1**	**41.7**

Semantic FPN output performs the best

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Overview



1. Which output from the FPN network to use for high resolution feature extraction?
2. **Why encode position information?**
3. The convolution step

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

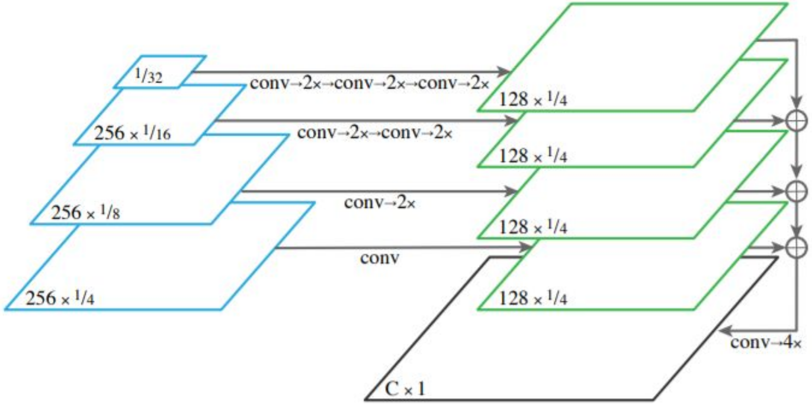
Feature Encoder

Why encode position information?

- We lose positional information due to multiple stages of upsampling and downsampling
- Encoding positional information brings better results

$coord_w$	$coord_f$	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✗	✗	39.9	45.0	32.4	29.9	41.2
✓	✗	39.9	45.0	32.2	30.0	41.1
✗	✓	40.2	45.3	32.5	30.4	41.6
✓	✓	**41.3**	**46.9**	**32.9**	**32.1**	**41.7**

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

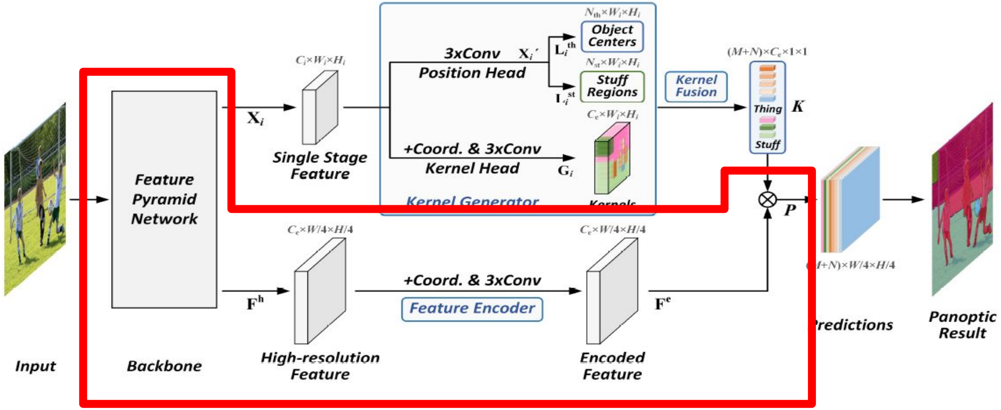


Semantic FPN

Figure_Source:Figure_Source:https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c

Feature Encoder

Overview

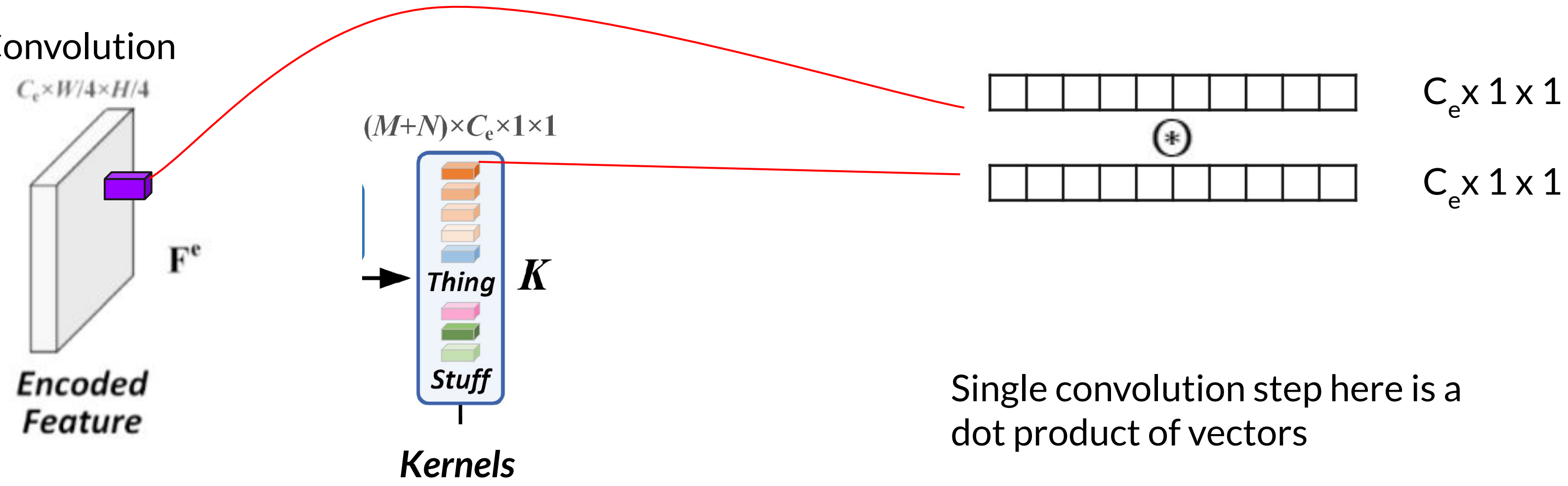


1. Which output from the FPN network to use for high resolution feature extraction?
2. Why encode position information?
3. **The convolution step**

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Convolution



Single convolution step here is a dot product of vectors

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Convolution

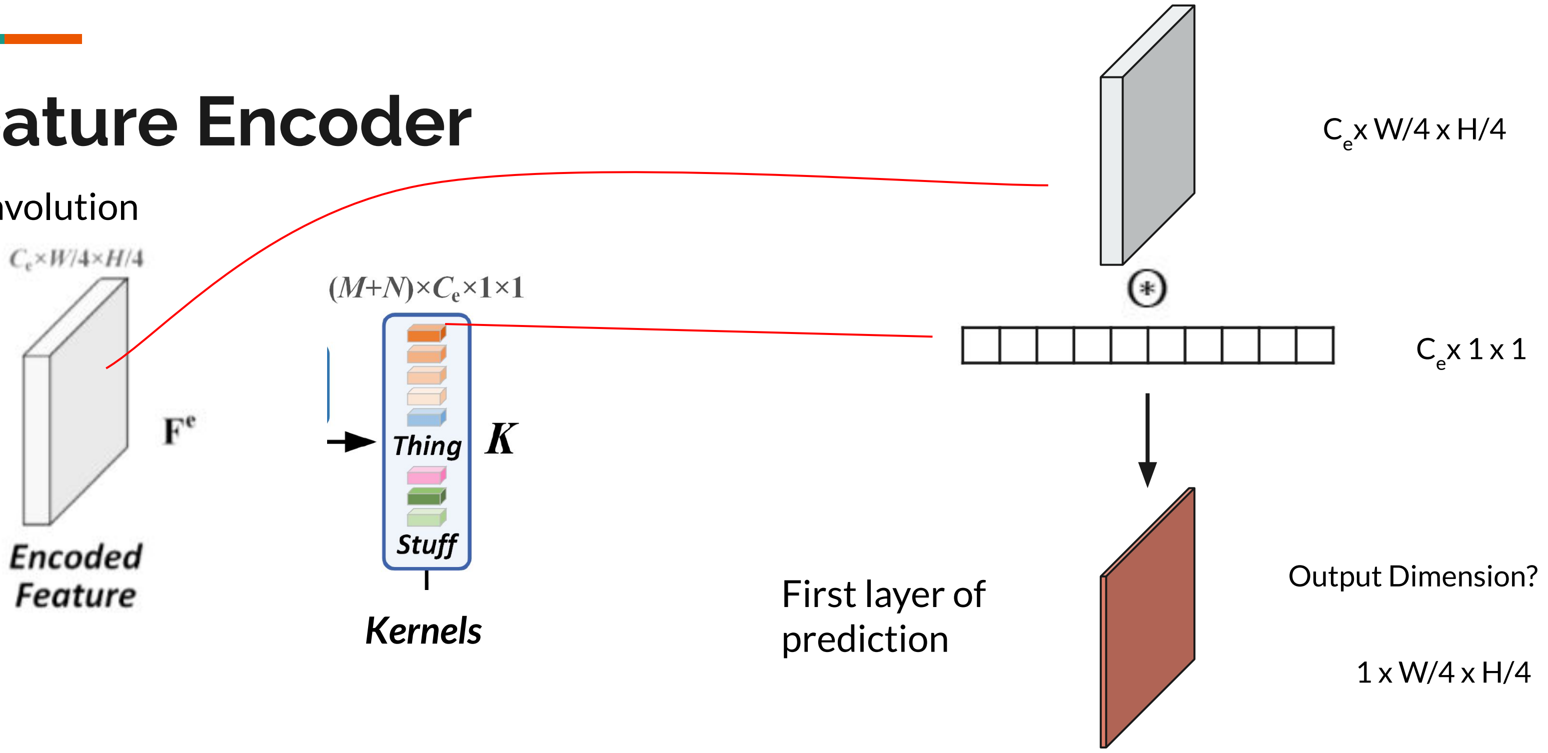


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Feature Encoder

Convolution

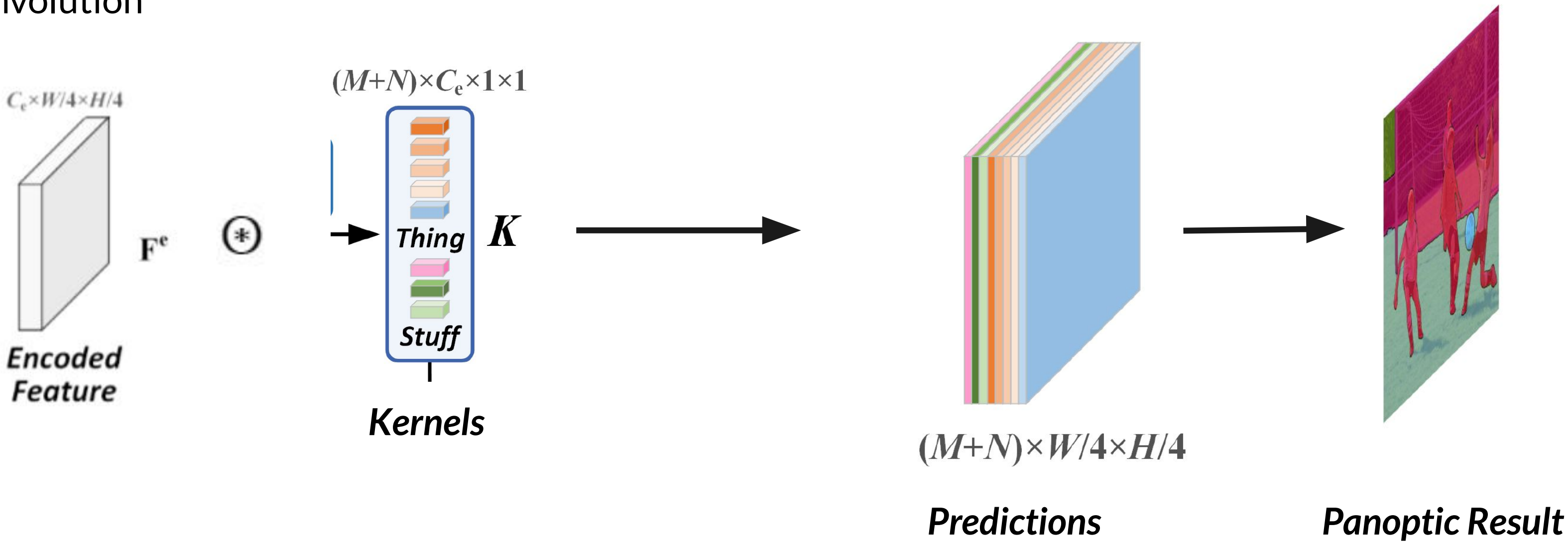


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Some Panoptic Segmentation Results - Visualization

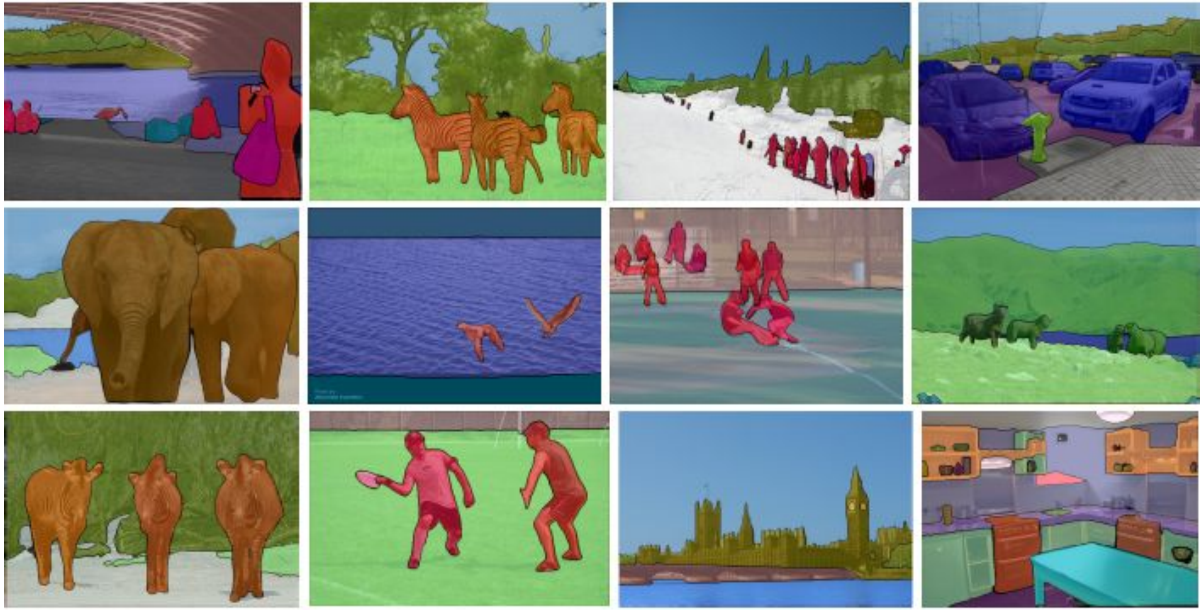


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Training and Inference - Training

Dice/F1 Loss - Recap

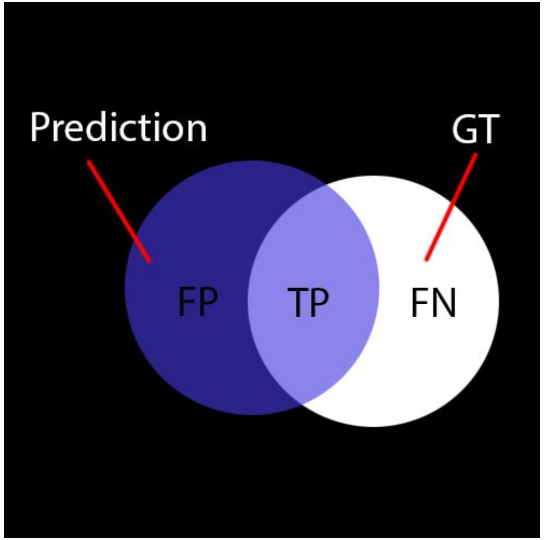
$$Dice\ score = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

$$Dice\ score = \frac{2\,TP}{2\,TP + FN + FP}$$

$$Dice\ score = \frac{2\,Intersection}{Intesection + Union}$$

$$Dice\ (For\ one\ instance) = \frac{2\,P_j\,Y_j^{seg}}{P_j + Y_j}$$

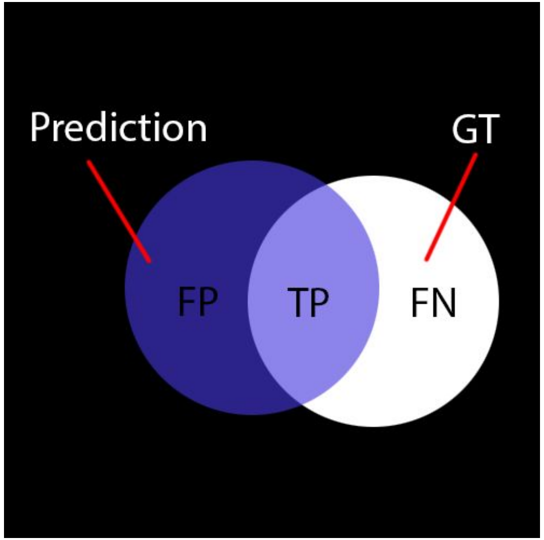
$$\boxed{Dice\ Loss = 1 - \frac{2\,P_j\,Y_j^{seg}}{P_j + Y_j}}$$



Training and Inference - Training

Dice/F1 Loss - Recap

$$Dice\ Loss = Dice(P_j, Y_j)$$



Training and Inference - Training

If there are M things and N stuff

Dice Loss

$$\mathcal{L}_{\text{seg}} = \sum_j \text{Dice}(\mathbf{P}_j, \mathbf{Y}_j^{\text{seg}})/(M+N),$$

To further release the potential of Kernel Generator - **Weighted Dice Loss**

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Training and Inference - Training

Weighted Dice Loss

- Multiple positives of each object is sampled.
- k positions of M things will be sampled in decreasing order of their scores

$$\mathrm{WDice}(\mathbf{P}_j, \mathbf{Y}_j^{\mathrm{seg}}) = \sum_k w_k \mathrm{Dice}(\mathbf{P}_{j,k}, \mathbf{Y}_j^{\mathrm{seg}}),$$

Where $w_k = s_k / \Sigma_i s_i$

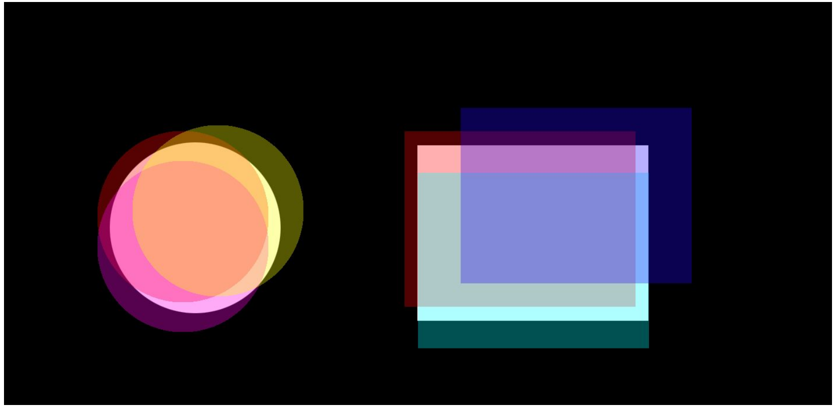


Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

Training and Inference - Training

Optimized target loss

$$\mathcal{L}_{\text{seg}} = \sum_j \text{WDice}(\mathbf{P}_j, \mathbf{Y}_j^{\text{seg}})/(M+N),$$

$$\mathcal{L} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}.$$

Figure from: Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., & Jia, J. (2021). Fully Convolutional Networks for Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 214-223).

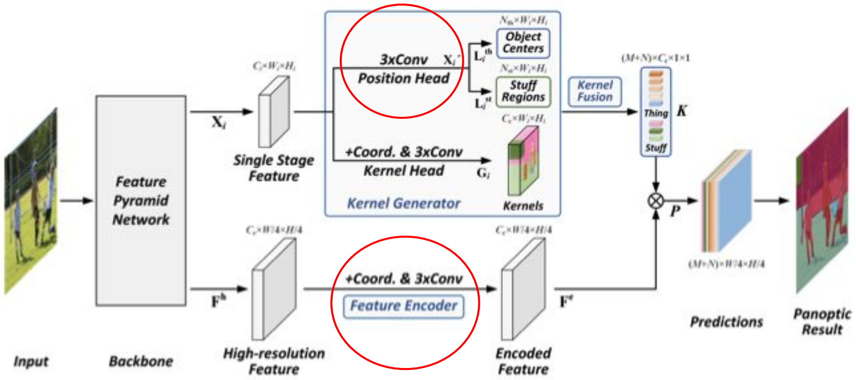
Optimization of the Model

How can the architecture be optimised to increase the panoptic quality?

Experiments

Experiments and Inference

deform	*conv num*	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✗	1	38.4	43.4	31.0	28.3	39.9
✗	2	38.9	44.1	31.1	28.9	40.1
✗	3	39.2	44.7	31.0	29.6	40.2
✗	4	39.2	44.9	30.8	29.4	39.9
✓	3	**39.9**	**45.0**	**32.4**	**29.9**	**41.2**



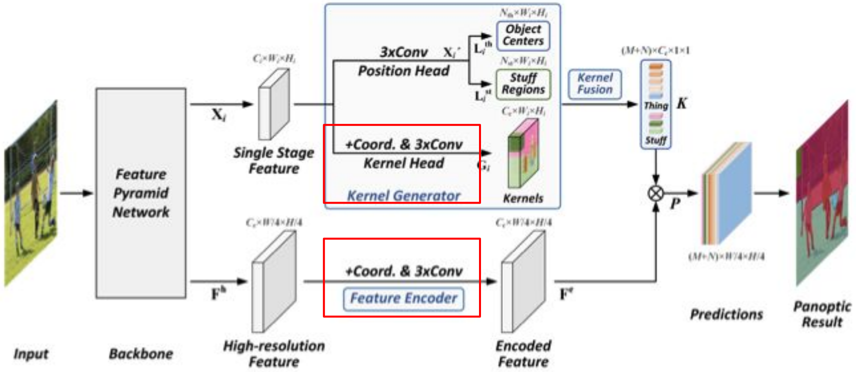
Achieves peak PQ at 3 stacked Conv3X3

Specifications of the model

# of convs						
3						

Experiments and Inference

$coord_w$	$coord_f$	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✗	✗	39.9	45.0	32.4	29.9	41.2
✓	✗	39.9	45.0	32.2	30.0	41.1
✗	✓	40.2	45.3	32.5	30.4	41.6
✓	✓	**41.3**	**46.9**	**32.9**	**32.1**	**41.7**



Specifications of the model

# of convs	Positional info encoding					
3	$Coord_{w,}$ $coord_f$					

Experiments and Inference

class-aware	*thres*	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✓	0.80	39.7	44.3	32.9	29.9	41.7
✓	0.85	40.8	46.1	32.9	31.5	41.7
✓	0.90	**41.3**	46.9	**32.9**	**32.1**	**41.7**
✓	0.95	41.3	**47.0**	32.9	31.1	41.7
✓	1.00	38.7	42.6	32.9	25.4	41.7
✗	0.90	41.2	46.7	32.9	30.9	41.7

The network attains the best performance with thres 0.90.

Specifications of the model

# of convs	**Combining coordinates**	**Threshold of Kernel Fusion**				
3	$Coord_{w,}$ $coord_f$	0.90				

Experiments and Inference

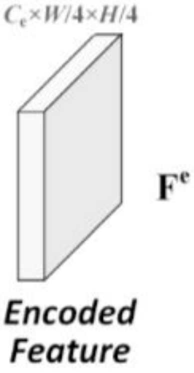
kernel-fusion	*nms*	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✗	✗	38.7	42.6	32.9	25.4	41.7
✗	✓	38.7	42.6	32.9	27.8	41.7
✓	✗	**41.3**	**46.9**	**32.9**	32.1	**41.7**
✓	✓	41.3	46.9	32.8	**32.3**	41.7

Specifications of the model

# of convs	**Combining coordinates**	**Threshold of Kernel Fusion**	**Method of removing repetitive predictions**			
3	$Coord_{w,}$ $coord_f$	0.90	Kernel Fusion only			

Experiments and Inference

channel num	PQ	PQ^{th}	PQ^{st}	AP	mIoU
16	39.9	45.0	32.1	30.8	41.3
32	40.8	46.3	32.5	31.7	41.6
64	**41.3**	46.9	**32.9**	32.1	**41.7**
128	41.3	**47.0**	32.6	**32.6**	41.7



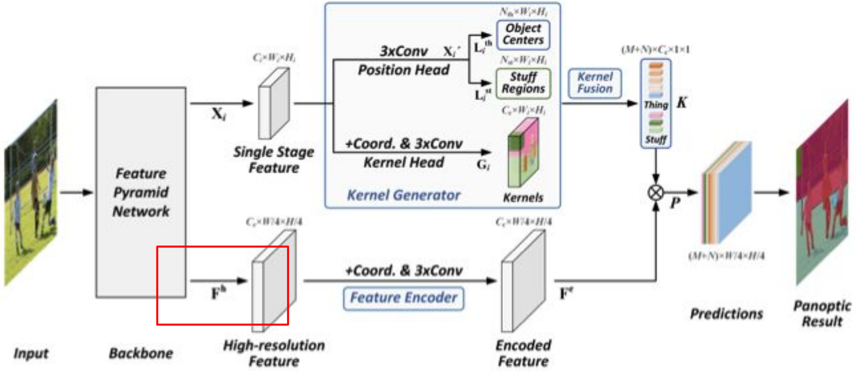
Highest PQ with 64 channels,
and extra channels contribute little improvement

Specifications of the model

# of convs	Combining coordinates	Threshold of Kernel Fusion	Method of removing repetitive predictions	# of channels		
3	$Coord_{w,}$ $coord_f$	0.90	Kernel Fusion only	64		

Experiments and Inference

feature type	PQ	PQ^{th}	PQ^{st}	AP	mIoU
FPN-P2	40.6	46.0	32.4	31.6	41.3
FPN-Summed	40.5	46.0	32.1	31.7	41.1
Semantic FPN [18]	**41.3**	**46.9**	**32.9**	**32.1**	**41.7**



Specifications of the model

# of convs	Combining coordinates	Threshold of Kernel Fusion	Method of removing repetitive predictions	# of channels	High res feature generator method	
3	Coord$_{w,}$ coord$_f$	0.90	Kernel Fusion only	64	Semantic FPN	

Experiments and Inference

weighted	k	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✗	-	40.2	45.5	32.4	31.0	41.3
✓	1	40.0	45.1	32.4	30.9	41.4
✓	3	41.0	46.4	32.7	31.6	41.4
✓	5	41.0	46.5	32.9	32.1	41.7
✓	7	**41.3**	**46.9**	**32.9**	**32.1**	41.7
✓	9	41.3	46.8	32.9	32.1	**41.8**

Best PQ with 7 top-scoring kernels

Specifications of the model

# of convs	Combining coordinates	Threshold of Kernel Fusion	Method of removing repetitive predictions	# of channels	High res feature generator method	K in weighted Dice Loss
3	$Coord_{w,}$ $coord_f$	0.90	Kernel Fusion only	64	Semantic FPN	7

Results for Panoptic FCN on COCO val-dev set

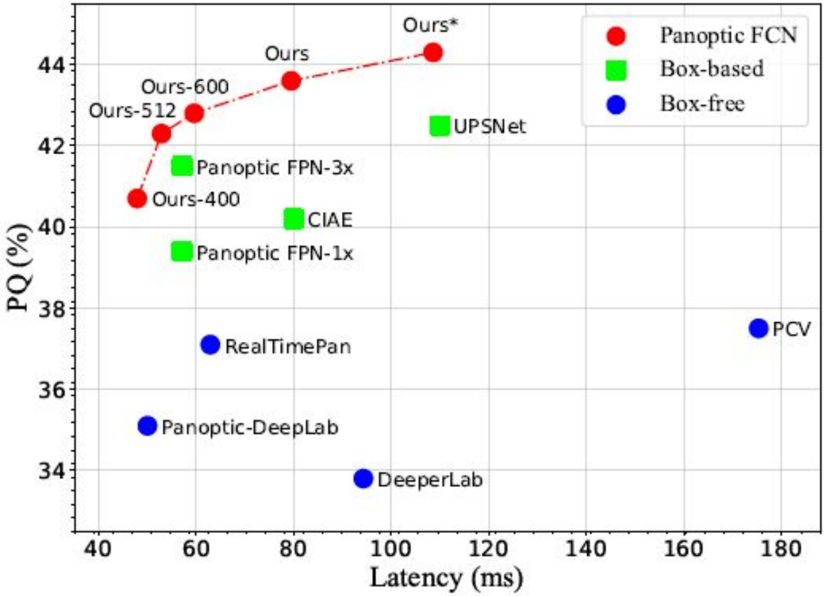
Method	Backbone	PQ	PQ^{th}	PQ^{st}
		box-based		
Panoptic FPN [18]	Res101-FPN	40.9	48.3	29.7
CIAE [11]	DCN101-FPN	44.5	49.7	36.8
AUNet [25]	ResNeXt152-FPN	46.5	**55.8**	32.5
UPSNet [50]	DCN101-FPN	46.6	53.2	36.7
Unifying‡ [24]	DCN101-FPN	47.2	53.5	37.7
BANet [5]	DCN101-FPN	47.3	54.9	35.9
		box-free		
DeeperLab [51]	Xception-71	34.3	37.5	29.6
SSAP [10]	Res101-FPN	36.9	40.1	32.0
PCV [43]	Res50-FPN	37.7	40.7	33.1
Panoptic-DeepLab [6]	Xception-71	39.7	43.9	33.2
AdaptIS [40]	ResNeXt-101	42.8	53.2	36.7
Axial-DeepLab [44]	Axial-ResNet-L	43.6	48.9	35.6
Panoptic FCN	Res101-FPN	45.5	51.4	36.4
Panoptic FCN	DCN101-FPN	47.0	53.0	37.8
Panoptic FCN*	DCN101-FPN	47.1	53.2	37.8
Panoptic FCN*‡	DCN101-FPN	**47.5**	53.7	**38.2**

Highest PQ value with the Enhanced Panoptic FCN model

Speed Accuracy Results

Speed Accuracy

Surpasses all previous models by a large margin in terms of speed-accuracy balance



Other Panoptic Segmentation Models

- Mask Former - July 2021
- Max-DeepLab - April 2021
- Panoptic Seg Former - September 2021

There is high momentum in the research of Panoptic Segmentation with the advent of Transformer Encoders!!

Thank You

Questions?