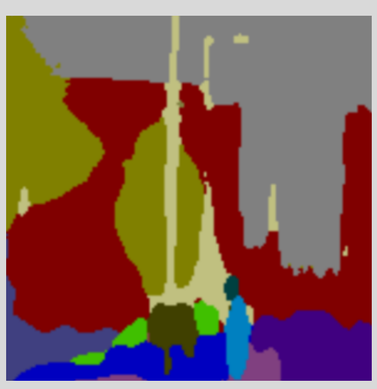
Feature-Guided Upsampling for Semantic Segmentation

Raymond Weiming Chan
MScAC, University of Toronto

Motivation

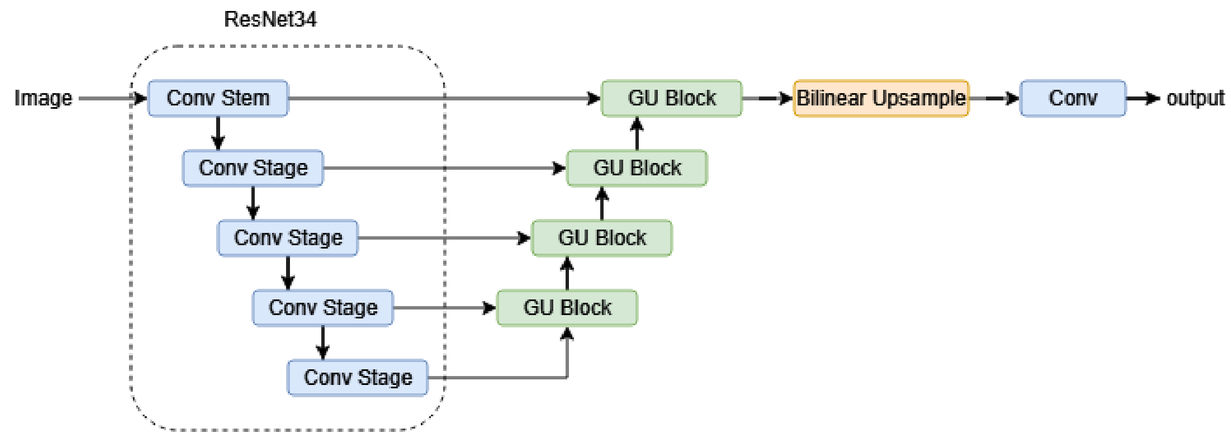
- Semantic segmentation models typically downsample images progressively to understand context. This process destroys fine spatial details.
- In the context of autonomous driving perception, it results in poor results near objects like poles and in areas with complex detail.



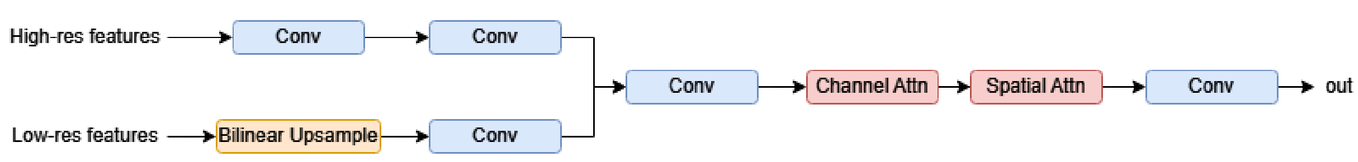
- Standard upsampling methods like bilinear interpolation ignore image content and cannot recover edges that were lost during downsampling.

New Technique

- We introduce a novel feature-guided upsampling network for semantic segmentation that utilizes the high-resolution features to guide the upsampling process.
- The architecture is a U-Net with a pretrained ResNet34 backbone, and a custom decoder that performs guided upsampling.



Our modified U-Net



Our GU Block (Guided Upsampling Block)

Related Work

- The encoder-decoder architecture is a popular choice in semantic segmentation. U-Net [1] uses transposed convolution, and DeepLab [2] uses bilinear interpolation. Both methods are not considered content-aware, and they fail to recover fine spatial details lost during encoding.
- PixelShuffle [3] is a learnable upsampling approach that uses sub-pixel convolution. Still, it lacks a guidance mechanism to exploit the spatial cues in the high-resolution input.
- Convolutional Block Attention Module (CBAM) [4] introduces the Channel Attention Module and the Spatial Attention Module, originally designed for classification and detection tasks.

References

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015.
[2] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.
[3] Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[4] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

Experimental Results

	mIoU	Pixel Accuracy	Dice Score
U-Net w/ Bilinear Upsampling Decoder	0.4489	0.9096	0.5485
Our Model	0.4578	0.9134	0.5646

- Our model achieves better results than baseline with the same training settings with minimal computational overhead.
- We have shown that our decoder architecture can be combined with pre-trained encoders to improve semantic segmentation
- Our experiment demonstrates that attention applied exclusively in the decoder's fusion blocks is a computationally efficient strategy that does not sacrifice performance