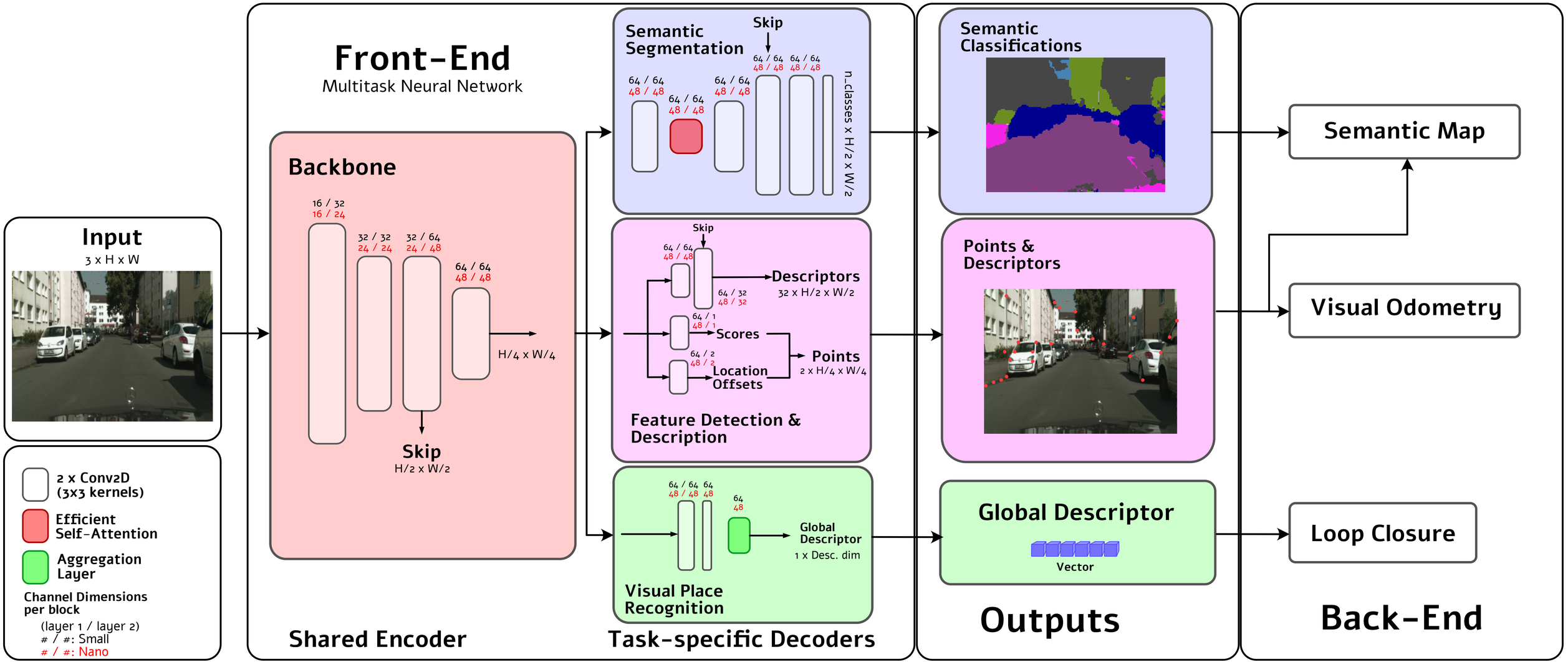
Front-End
Multitask Neural Network
Input
3 x H x W
Backbone
16 / 32
16 / 24
32 / 32
24 / 24
32 / 64
24 / 48
64 / 64
48 / 48
H/4 x W/4
Skip
H/2 x W/2
Semantic Segmentation
Skip
64 / 64
48 / 48
n_classes x H/2 x W/2
Descriptors
32 x H/2 x W/2
64 / 32
48 / 32
64 / 1
48 / 1
Scores
64 / 2
48 / 2
Location Offsets
Points
2 x H/4 x W/4
Feature Detection & Description
64 / 64 64
48 / 48 48
64
48
Global Descriptor
1 x Desc. dim
Visual Place Recognition
Semantic Classifications
Points & Descriptors
Global Descriptor
Vector
Semantic Map
Visual Odometry
Loop Closure
2 x Conv2D (3x3 kernels)
Efficient Self-Attention
Aggregation Layer
Channel Dimensions per block
(layer 1 / layer 2)
/ #: Small
/ #: Nano
Shared Encoder
Task-specific Decoders
Outputs
Back-End