Entire Scene

I

I_C

Cropped Image

Mask M

Global Feature Extractor (GFE)

Local Feature Extractor (LFE)

Intrinsic Feature Extractor

F_G

F_L

F_I

δ_G $\mathbf{f}_G$

δ_L $\mathbf{f}_L$

δ_I $\mathbf{f}_I$

Interpreter $\mathcal{I}$

Complete Features $\mathbf{f}_T \in \mathbb{R}^l$

Attributes vector $\hat{\mathbf{a}} \in \mathbb{R}^{n_c}$

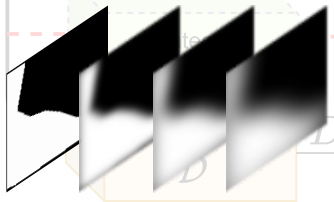
A_G $\mathcal{G}_G$

A_L $\mathcal{G}_L$

A_I $\mathcal{G}_I$

Category Embedding (e.g. floor)

$\hat{\mathbf{c}}$



Using Informed Convolution Layers

Concatenation

δ Feature-level Multiplication & Spatial Reduction

$\mathcal{G}$ Attention Gate

$\hat{\mathbf{c}}$ Estimated from LFE by Category Estimator