Graph-Boosted Attentive Network for Semantic Body Parsing

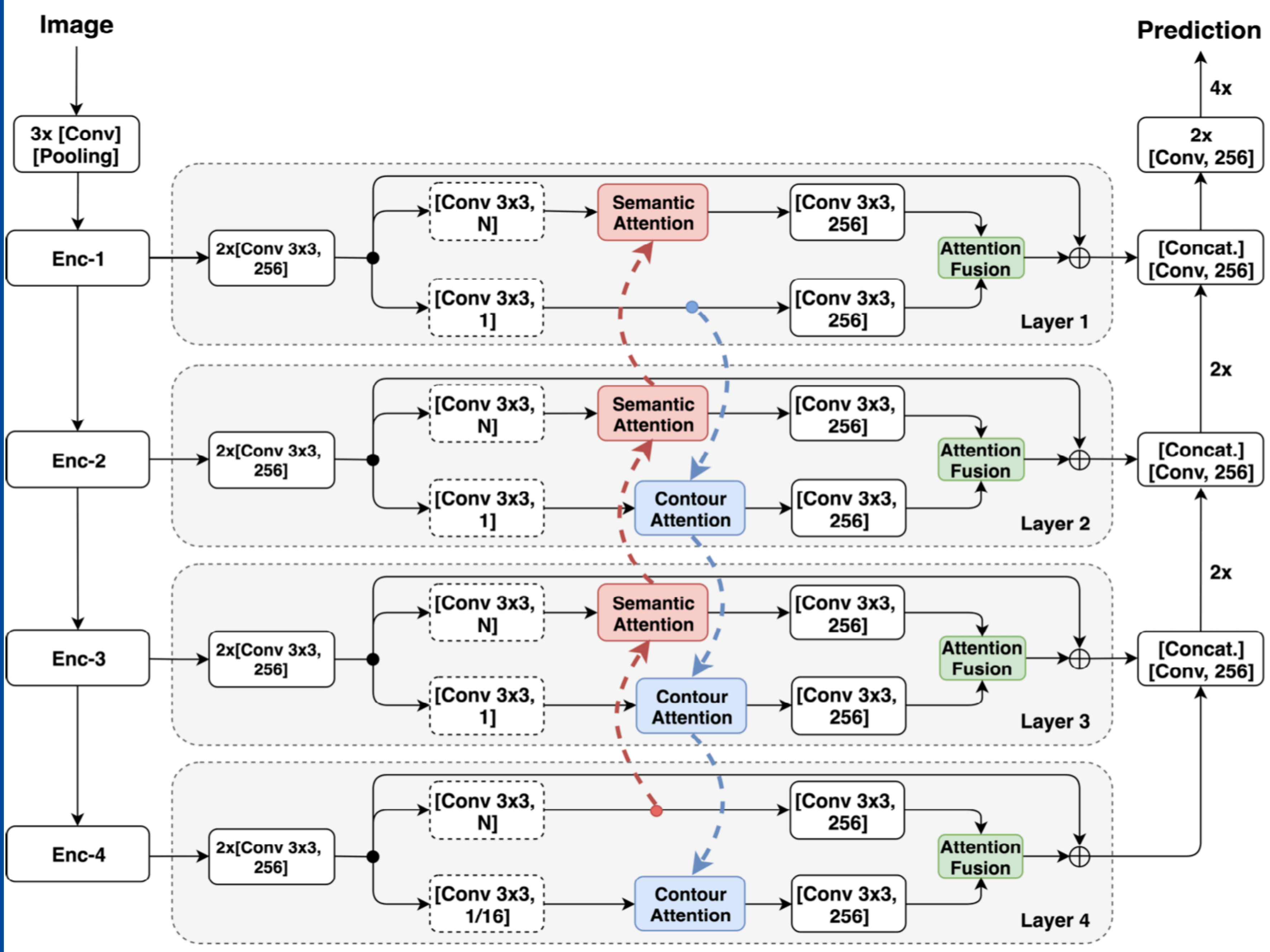
Tinghuai Wang
Nokia Technologies, Finland

Huiling Wang
Tampere University, Finland

Introduction

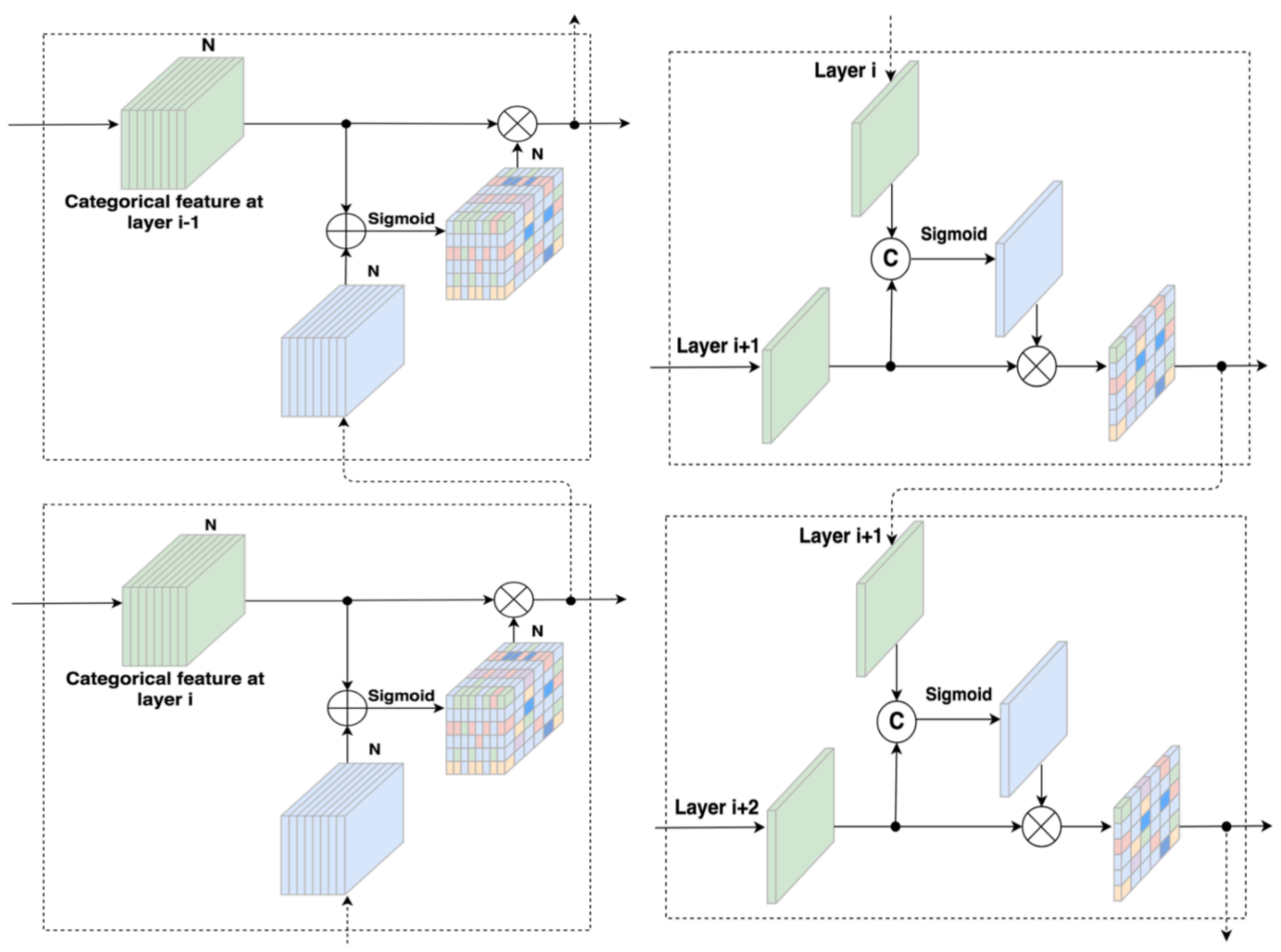
This paper proposes a novel approach to decomposing multiple human bodies into semantic part regions in unconstrained environments. Specifically we propose a convolutional neural network (CNN) architecture which comprises of novel semantic and contour attention mechanisms across feature hierarchy to resolve the semantic ambiguities and boundary localization issues related to semantic body parsing. We further propose to encode estimated pose as higher-level contextual information which is combined with local semantic cues in a novel graphical model in a principled manner.

Proposed CNN Architecture



Proposed Attention Modules

We propose novel semantic attention (left) and contour attention (right) modules:



Our attention mechanism is element-wise attention which is different from the channel-wise attention model. Our element-wise attention model suits better for semantic segmentation which is a dense prediction problem.

Overall, there are consistent semantic and contour information flows across the CNN feature hierarchy which are missing in the state-of-the-art architectures.

Pose as Context

We extract human skeleton map using Deeper-Cut to generate multi-layer superpixels with different granularities. In each superpixel map, we compute the geodesic distance from all superpixels w.r.t. the set of superpixels associated with each skeleton line. Based on the geodesic distance, the likelihood that semantic part Θp occurs at superpixel yi can be computed as

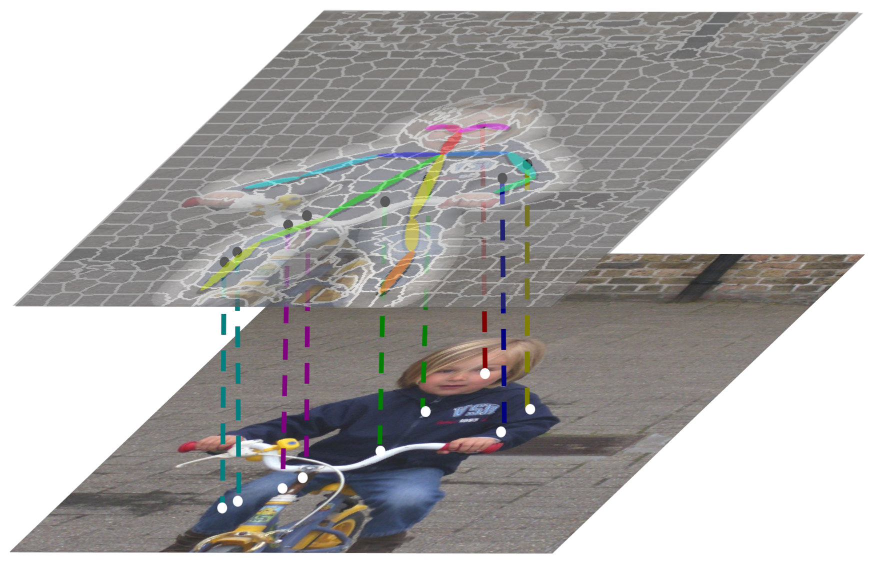
$$p(y_i|\Theta_p) = \exp(-\beta d_{geo}^2(y_i, \Theta_p))$$

Graphical Model

We construct an undirected graph G = (V, E) with pixels and superpixels as nodes V ={X, Y} respectively. Pose-Pixel Edge E_{XY} connects each superpixel and its constituent pixels; all spatially adjacent pixels are connected to form pixel edges E_{XX}; all spatially adjacent superpixels are connected to form superpixel edges E_{YY}. The cost functions are defined as

$$\begin{aligned} J_l^{\mathrm{X}} =& J_{l,U}^{\mathrm{X}} + J_{l,P}^{\mathrm{X}} + J_{l,C}^{\mathrm{X}} \\ =& \sum_{i=1}^{N_{\mathrm{X}}} \lambda^{\mathrm{X}} d_i^{\mathrm{X}} (u_{il} - \tilde{u}_{il})^2 + \sum_{i,j=1}^{N_{\mathrm{X}}} w_{ij}^{\mathrm{XX}} (u_{il} - u_{jl})^2 \\ &+ \sum_{i=1}^{N_{\mathrm{X}}} \pi d_i^{\mathrm{X}} (u_{il} - \bar{u}_{il})^2 \end{aligned}$$

$$\begin{aligned} J_l^{\mathrm{Y}} =& J_{l,U}^{\mathrm{Y}} + J_{l,P}^{\mathrm{Y}} + J_{l,C}^{\mathrm{Y}} \\ =& \sum_{m=1}^{N_{\mathrm{Y}}} \lambda^{\mathrm{Y}} d_m^{\mathrm{Y}} (v_{ml} - \tilde{v}_{ml})^2 + \sum_{m,n=1}^{N_{\mathrm{Y}}} w_{mn}^{\mathrm{YY}} (v_{ml} - v_{nl})^2 \\ &+ \sum_{m=1}^{N_{\mathrm{Y}}} \psi d_m^{\mathrm{Y}} (v_{ml} - \bar{v}_{ml})^2 \end{aligned}$$



Posterior probabilities of each pixel with respect to part label l can then be computed following Bayes rule

$$p(l|x_i) = \frac{p(x_i|l)p(l)}{\sum_{l'=1}^{L} p(x_i|l')p(l')} = \frac{u_{il}}{\sum_{l'=1}^{L} u_{il'}}$$

Each pixel is finally assigned with the label corresponding to the class with the maximum a posterior probability.

Results

Table 1: Quantitatively segmentation results on Pascal Person-Part dataset

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Avg.
DeepLab-LargeFOV [3]	78.09	54.02	37.29	36.85	33.73	29.61	92.85	51.78
HAZN [37]	80.79	59.11	43.05	42.76	38.99	34.46	93.59	56.11
Attention [5]	-	-	-	-	-	-	-	56.39
LG-LSTM [19]	82.72	60.99	45.40	47.76	42.33	37.96	88.63	57.97
Graph LSTM [18]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.16
DeepLab v2 [4]	-	-	-	-	-	-	-	58.90
JPS (final, CRF) [38]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
PCNet-126 [43]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Our model (w/o graph)	89.19	74.88	55.98	60.76	50.76	41.45	95.12	66.87
Our model (final)	90.84	75.85	56.18	64.86	52.86	43.52	95.75	**68.55**