This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Cross-Granularity Attention Network for Semantic Segmentation

Lingyu Zhu[1*] Tinghuai Wang[2*] Emre Aksu[2] Joni-Kristian Kämäräinen[1]
[1] Tampere University, Finland
[2] Nokia Technologies, Finland

## Abstract

*Despite the remarkable progress of semantic segmentation in recent years, much remains to be addressed in order to achieve better semantic coherence and boundary delineation. In this paper, we propose a novel convolutional neural network (CNN) architecture for semantic segmentation which explicitly addresses these two issues. Specifically, we propose a categorical attention mechanism to propagate consistent category-oriented information across multi-granularity contextual interpretations to close the semantic gap residing in CNN feature hierarchy. This novel design alleviates the semantic information loss during the feature combination and transformation process in decoder network. We further integrate a contour branch in our architecture to enhance the boundary awareness of the semantic feature derived in the form of a novel element-wise contour attention at each level of feature hierarchy. Additionally, we introduce a cross-granularity contour enhancement mechanism to propagate rich boundary cues from early layers to deep layers. We perform extensive quantitative evaluations in close proximity to object boundaries which confirms its superior effectiveness in boundary delineation. These novel mechanisms which boost the essentials in segmentation, i.e., region-wise semantic coherence and accurate object contour localization, allow our architecture "MeshNet" to obtain state-of-the-art performance on two challenging datasets, i.e., PASCAL VOC 2012 and Cityscapes.*

## 1. Introduction

Recently, semantic image segmentation has achieved significant improvements in accuracy by utilizing convolutional neural networks (CNNs) [38] due to rich information of object categories and scene semantics learned from diverse set of images.

However, the state-of-the-art CNN architectures are still challenged by semantic ambiguities and poor boundary delineation as shown in Figure 1. The former problem, *i.e.* semantic ambiguities, is mainly caused by the semantic gap between feature hierarchies of CNN layers, where earlier layers are lacking sufficient semantic knowledge to make accurate semantic labelling based on local features, despite

(a) Image (b) DeepLabv3+ (c) MeshNet

Figure 1: Comparison on images with semantically ambiguous object (*i.e.* table in the first row) and thin structures (*i.e.* bicycle in the second row). The state-of-the-art encoder-decoder architecture DeepLabv3+ [15] (the second column) fails to either recognize the partially occluded table or delineate the boundaries of bicycle, whereas the proposed architecture "MeshNet" (the third column) excels in both situations.

of their high spatial resolution. Albeit this issue is alleviated to some extent by the adoption of skip connections in encoder-decoder architectures [38, 42, 2, 15], repeatedly merging features with lower-level features of earlier layers inevitably dilutes semantic information. This is also the root of the poor boundary delineation issue since the boundary information is mainly preserved in earlier layers that consequently gets smoothed out by the coarse feature map from deeper layers during the iterative combination and transformation process. More importantly, this extracted boundary related information is isolated from the semantic information which might harm the intra-class homogeneity during dense prediction. This poor boundary awareness might in turn deteriorate the semantic prediction as psychophysical studies [5, 50] show that human beings can recognize objects using fragments of outline contour alone.

Motivated by the above, in this paper we design a novel deep encoder-decoder architecture to principally address

*Equal contribution; this work is done in Nokia Technologies

both the semantic gap and boundary diminishing problems in semantic segmentation with three main contributions. Firstly, we explicitly bridge the semantic gap between feature hierarchies by proposing a cross-granularity categorical attention mechanism. Leveraging deep supervision, our categorical attention module is able to enforce the feature adaptation at each hierarchy to follow the consistent top-down categorical attention, selecting the most category-relevant feature across the spatial granularities. This novel cross-granularity categorical attention mechanism alleviates the semantic information loss during the feature combination and transformation process. Secondly, we embed contour detection in a novel form of element-wise contour attention at each layer of feature hierarchy to explicitly integrate contour information to enforce semantic regions to obey discriminative visual features in the image. Thirdly, we introduce a cross-granularity contour enhancement mechanism to propagate the rich boundary cues from shallower layers to deeper layers. To the best of our knowledge, we are the first to introduce multiple contour detection networks with accordance to CNN feature hierarchy to explicitly extract boundary information and propose cross layer propagation in order to resolve the long standing boundary diminishing problem.

## 2. Related Work

Significant improvement in semantic segmentation has been witnessed since the development of Fully Convolutional Networks (FCNs) [44, 38]. Various FCN based architectures [19, 17, 7, 66, 33, 15, 60, 62, 6, 64, 61, 27, 63] have been proposed to prevalently exploit contextual information from feature pyramid or attention mechanism. Yet, previous works addressing the semantic gap and boundary awareness issues in FCNs are sparse.

### Spatial pyramid pooling:

Diverse range of contextual information is playing an important role in capturing finer feature delineation and is widely employed in different semantic segmentation tasks. PSPNet [65], for instance, captures and aggregates features from multiple receptive fields. In addition, DeepLabv2 [12] introduces Atrous Spatial Pyramid Pooling (ASPP) to combine multi-scale features from parallel atrous convolution layers with different dilation rates. Recently, DenseASPP [60] utilizes densely connected atrous convolutions to generate large scale range features densely.

### Encoder-decoder:

FCNs naturally encode multi-scale contextual information in different levels of features. Encoder-decoder architectures [2, 42, 15] have been proposed to integrate feature hierarchies from encoder to refine the final prediction. This line of work is mainly motivated by the need of recovering the reduced spatial information of CNN caused by strided convolution and pooling operations. For example, DeconvNet [41] employs deep deconvolutions and unpooling layers to construct the final semantic segmentation result. U-net [42] has a contracting path and a symmetric expanding path with skip connections between each encoder and corresponding decoder layer. RefineNet [33] exploits features with a multi-path refinement network in a recursive manner. Bilinski and Prisacariu [6] add dense shortcut connections from feature hierarchy to merge semantic feature maps from all previous decoder levels. Most recently, DeepLabv3+ [15] adopts a simple decoder with one skip connection from low stage to recover the object boundaries.

### Attention mechanism:

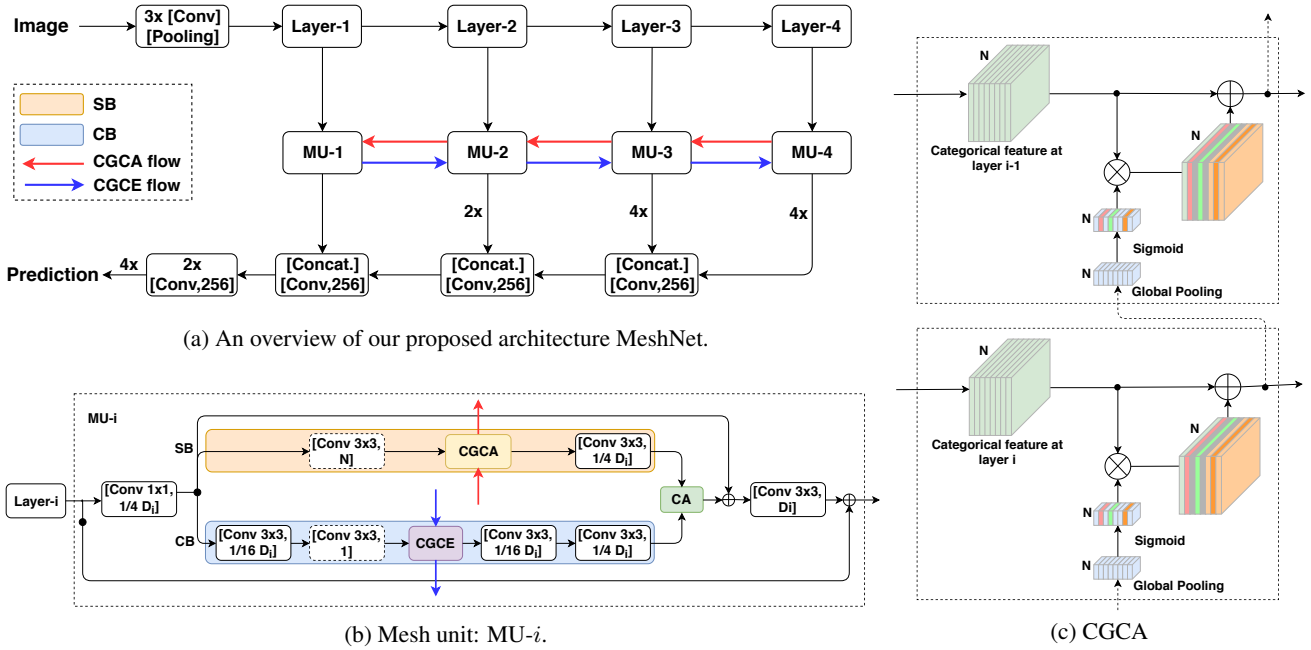
Recent studies have shown the gains of introducing attention insights into different structural prediction tasks. From image classification [57, 9, 24, 40], localization [8, 1], visual captioning [59, 29], visual question answering [10] to natural language processing [3]. For semantic segmentation, Chen *et al*. [14] proposes an attention module to softly weight multi-scale features. SENet [27] exploits channel dependences by squeezing global spacial information into channel-wise statistics. EncNet [63] selectively highlights class-dependent feature maps and integrates global context information by a separate encoding layer. DFN [62] utilizes global average pooling to introduce channel-wise attention into network concerning the selection of more discriminative features. Inspired by the pioneering work, we propose a cross-granularity categorical attention mechanism to bridge the semantic gap between the feature maps of deeper layers and shallower layers; we also introduce a novel cross-granularity contour enhancement mechanism to convey rich boundary information from lower hierarchy to higher hierarchy, and an element-wise contour attention module to explicitly enhance the semantic boundary awareness at each feature hierarchy.

### Contour detection:

Although some recent CNN based contour detectors [4, 45, 28, 58] have been proposed, previous works explicitly incorporating contour feature to improve semantic segmentation in an end-to-end manner are sparse in the literature. Lately, DFN [62] combines a decoupled border prediction network in parallel with its proposed segmentation decoder network to adapt the early feature representation from the encoder network with respect to semantic boundaries. The boundary information is not effectively utilized to enahance the feature representations. On the contrary, our architecture directly embeds a contour detection network in accordance to each layer of the feature hierarchy and introduces an element-wise contour attention module to explicitly enhance boundary awareness of semantic features.

## 3. Methods

Our MeshNet architecture consists of two main parts: an encoder and a decoder, as illustrated in Figure 2. The encoder comprises of four layers according to the size of feature maps, namely Layer-1, -2, -3 and -4 respectively, whilst the rest of the blocks constitute the decoder. Encoder extracts appearance and contextual information at various




(a) An overview of our proposed architecture MeshNet.

(b) Mesh unit: MU-$i$.

(c) CGCA

Figure 2: (a) An overview of our proposed architecture MeshNet. Four mesh units MU-$i$ take as input of the feature maps from encoder, which propagate consistent categorical information (red arrows) in a top-down manner across feature hierarchy and explicitly enhance boundary awareness of semantic features by incorporating contour features (blue arrows) from early hierarchy. (b) Architecture of MU-$i$. In each MU-$i$, the feature maps from Layer-$i$ are fed into two branches, *i.e.*, semantic branch (SB) and contour branch (CB). The features from two branches are fused by an element-wise Contour Attention (CA) module to explicitly enhance the semantic boundary awareness at each feature hierarchy. The dashed blocks in each MU-$i$ are associated with the auxiliary loss functions defined in Section 3.5 during training. (c) Architecture of proposed CGCA module. CGCA: Cross-Granularity Categorical Attention, and CGCE: Cross-Granularity Contour Enhancement.

hierarchies, with decreasing spatial details and increasing semantic information from Layer-1, -2, -3 to Layer-4. At each layer Layer-$i$ ($i = 1, 2, 3, 4$), the encoder interfaces with the proposed decoder by providing the feature map as input to its corresponding *Mesh Unit* (MU), *i.e.* MU-$i$.

### 3.1. Mesh Unit

Denoting the feature dimension from Layer-$i$ as $D_i$, the corresponding mesh unit MU-$i$ firstly reduces the feature dimension to $D_i/4$, using either an $1\times1$ convolutional layer (for MU-1, -2 and -3) or an Atrous Spatial Pyramid Pooling (ASPP) [13] (for MU-4). Thereafter, the feature map goes though a $3\times3$ convolution layer with $N$ neurons, where $N$ is the number of categories, whereby the feature map is explicitly projected into the *categorical feature* space with category-wise deep supervisions (see Section 3.5). Thereby, the redundancies of features, which are irrelevant with respect to semantic prediction, are suppressed, leaving compact and essential features encoding more "focused" categorical information.

This *categorical feature* is forwarded to the proposed *cross-granularity categorical attention* module (Section 3.2) for enhancing the categorical information with adjacent mesh units, which is consecutively projected back to $D_i/4$ dimensions and fused with the feature map from *contour branch* in the proposed *contour attention* module (Section 3.3). The fused feature is thereafter projected back to $D_i$. In order to enable fast convergence and avoid feature degradation, two residual connections [26] are added to each mesh unit whereby $D_i$ and $D_i/4$ feature maps are summed respectively. The feature map is then gradually projected to output space via $3\times3$ convolution layers after concatenation with features from deeper mesh unit.

### 3.2. Cross-Granularity Categorical Attention

Most modern semantic segmentation networks neglect the semantic ambiguity issue caused by the semantic gap between feature hierarchies from CNN, where deeper layers encodes rich semantic information while earlier layers are lacking sufficient semantic knowledge to make accurate semantic labelling despite of their high spatial resolution. Recent encoder-decoder architectures attempt to address this issue by means of skip connections, whereas repeatedly integrating features across granularities dilutes semantic information and decreases inter-class distinction. We propose cross-granularity categorical attention (CGCA) module aiming to guide the category-oriented information propagation across multi-granularity structural interpretations as illustrated in Figure 2c.

In each CGCA, global average pooling followed by a

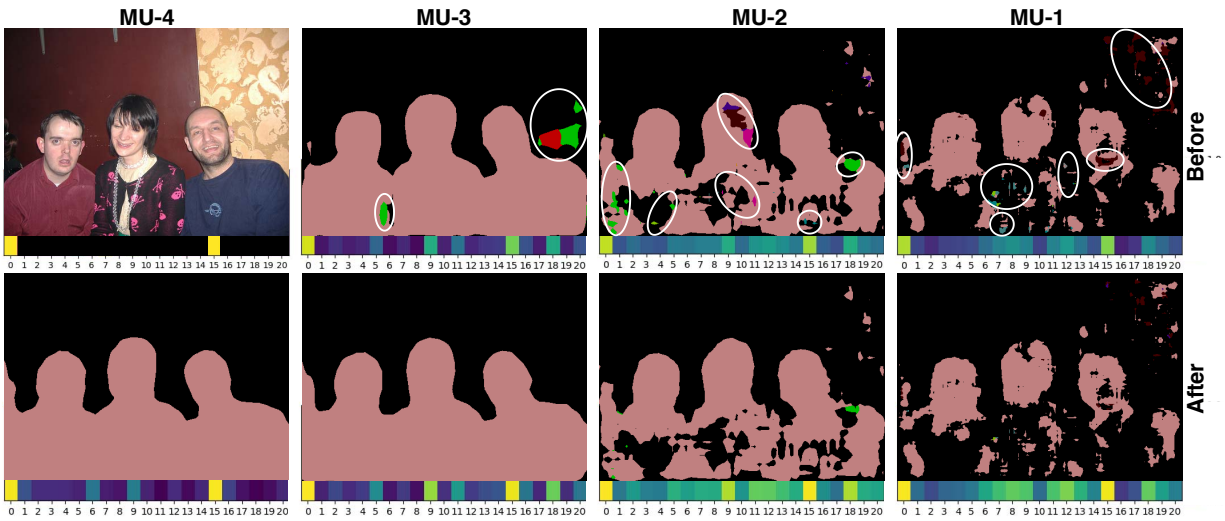



Figure 3: Intermediate semantic predictions before (top row) and after (bottom row) CGCA. The corresponding attention weight vector below each label map demonstrates the efficacy of the proposed architecture where the semantic ambiguity issue residing in the lower-level features is effectively resolved and the categorical feature becomes more attentive on consistent categories, *i.e*. background (0) and people (15) in this example.

sigmoid function is applied to the $N$-channel categorical feature map from higher layer (or mesh unit), which extracts the essence of categorical information, *i.e*. the global categorical attention. CGCA is achieved by multiplying the global categorical attention, *i.e*. the weight vector, with the current feature map to adjust channel-wise responses with respect to the predicted categories informed by higher layers. Finally, different from DFN [62], the category-enhanced feature is summed with the current lower-level feature map in order to preserve current hierarchical features. This attention plays a crucial role in maintaining consistent categorical information across feature hierarchy and bridging the semantic gaps. Figure 3 shows the intermediate semantic labellings as well as the categorical attentions before and after the CGCA module, where we can see that CGCA significantly resolves the semantic ambiguity issue residing in the lower-level features and renders the features more attentive on consistent categories, *i.e*. background (0) and people (15) in this example.

### 3.3. Contour Branch and Contour Attention

Contour features are capable of localizing objects in space and scale which in turn provide better boundary delineation and shape context cues for semantic segmentation task against within-class variations [52, 54, 46, 47]. Existing FCN based segmentation networks largely pay little attention to the boundary awareness due to its inherent design limitations, *i.e*. the boundary information is mainly preserved in earlier layers and its extraction is isolated from the semantic information residing in deeper layers. In order to incorporate object contour information at all hierarchical feature layers, and shift the contour information seamlessly to the segmentation task, we propose a simple yet effective contour branch (CB) and contour attention (CA) modules.

As illustrated in Figure 2(b), the contour branch consists of four convolutional layers with kernel size 3×3 and channel dimensions $D_i/16$, 1, $D_i/16$, and $D_i/4$ respectively. The contour prediction is trained with the supervision of boundaries that are generated by simply adopting Sobel edge detection on the segmentation ground-truth data whose loss is defined in Section 3.5. Note that, accurate contour prediction from the proposed contour branch is neither expected nor required since our ultimate goal is semantic segmentation and fragmental contour feature is sufficient to strengthen the boundary awareness and intra-class homogeneity of features.

Given the features of contour branch, we further propose a novel element-wise contour attention module to seamlessly integrate the learned boundary information with the semantic-rich features of segmentation task. As illustrated in Figure 4, the feature map with contour information from CB firstly goes through a sigmoid function without global pooling and then enhances the semantic features by an element-wise multiplication and a summation with the semantic branch features.

### 3.4. Cross-Granularity Contour Enhancement

Deeper layers usually have poor boundary delineation issue since the contour information is mainly preserved in shallow layers that consequently gets smoothed out while the model going deeper. In order to propagate the rich contour information from early layer to each deeper layer, we propose an effective cross-granularity contour enhancement (CGCE) module. CGCE achieves contour enhancement by simple element-wise summation between the learned contour information from early layer and the contour detection

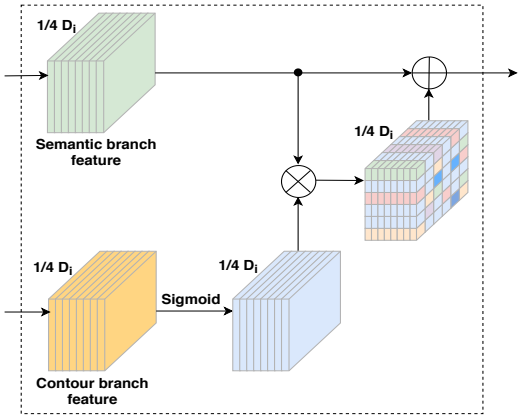



Figure 4: The architecture of contour attention module (CA). CA seamlessly integrate the learned boundary information from CB with the semantic-rich features of SB by passing CB features through a sigmoid function and then enhancing the semantic features by an element-wise multiplication and a summation with the SB features. CA: Contour Attention, CB: Contour Branch, and SB: Semantic Branch.

of current layer. Figure 5 shows the element-wise contour enhancement of four mesh units respectively, which demonstrates that the contour branch from each mesh unit is able to capture object boundaries despite of its simplicity and the proposed CGCE module significantly enhances the boundary awareness especially the deeper layers, *e.g*. MU-4.

### 3.5. Deep Supervision

To facilitate the learning of categorical and contour features w.r.t. CGCA and CGCE modules respectively, deep supervisions [31, 48] are adopted for both branches of each mesh unit. We denote the training dataset by $T = \{(X_k, Y_k), k = 1, 2, 3, ..., K\}$, where $Y_k = \{y_1^k, y_2^k, y_3^k, ..., y_{P_{X_k}}^k\}$ denotes the pixel-wise ground truth of the raw input image sample $X_k = \{x_1^k, x_2^k, x_3^k, ..., x_{P_{X_k}}^k\}$, $P_{X_k}$ is the total pixel counts of input image $X_k$, and $K$ is the number of training set samples. The value of $y_j^k$ is in the range of $\{0, 1, 2, ..., N-1\}$, and $N$ is the total number of categories of the training dataset. Denoted as $Z_k = \{z_1^k, z_2^k, z_3^k, ..., z_j^k\}$ with $z_j^k \in \{0, 1\}$, the object boundary ground-truth is produced by running a Sobel operator on each segmentation ground-truth. In addition to the principal softmax loss used to supervise the output of the whole MeshNet, two auxiliary loss functions are added to each mesh units — one to supervise the learning of categorical features, and the other to supervise contour feature learning. For simplicity, we denote the set of corresponding network weights as $\mathcal{W} = \{W_{(s,0)}, W_{(s,i)}, W_{(c,i)}\}$, where $s$ and $c$ represent semantic and contour predictions respectively, and $i$ ($i \in \{0, 1, 2, 3, 4\}$) refers to either the principal prediction (0) or auxiliary mesh unit predictions (1-4). Thereby, the final weighted loss function is defined as,

$$\begin{aligned}\mathcal{L}(\mathcal{W}, X_k, Y_k, Z_k) &= \ell_s(W_{(s,0)}, X_k, Y_k) \\ &+ \sum_{i=1}^{4} \alpha_{(s,i)} \ell_s(W_{(s,i)}, X_k, Y_k) \\ &+ \sum_{i=1}^{4} \alpha_{(c,i)} \ell_c(W_{(c,i)}, X_k, Z_k)\end{aligned} \tag{1}$$

where $\ell_s$ represents the softmax loss function for semantic prediction task, $\ell_c$ denotes the adaptive mini-batch weighted loss function for contour prediction tasks, $\alpha$ is the weight for each loss, which decides the contribution of each loss and $\alpha_{(s,i)} = \alpha_{(c,i)} = 0.025$ are empirically chosen for our training.

Due to the highly imbalanced distribution of the contour and non-contour pixels of most natural images, we adopt an adaptive loss balancing weight [58] $\beta = B_{Z_-}/B_Z$ and $1 - \beta = B_{Z_+}/B_Z$ to trade off recall and precision by increasing and decreasing the cost, where $B_{Z_+}$ and $B_{Z_-}$ denote the total pixels number of contour and non-contour region from the ground-truth labels of each mini batch, respectively. Specifically, the weighted cross-entropy loss function $\ell_c$ for contour loss is defined as,

$$\begin{aligned}\ell_c(w_{(c,i)}, X_k, Z_k) = &-\beta \sum_{j \in Z^k} z_j^k \log(p(w_{(c,i)}, x_j^k)) \\ &-(1-\beta) \sum_{j \in Z^k} (1 - z_j^k)(\log(1 - p(w_{(c,i)}, x_j^k)))\end{aligned} \tag{2}$$

where $p((w_{(c,i)}, x_j^k))$ is the contour prediction probability that calculated by a sigmoid function.

### 3.6. Encoder

We use Xception65 [16] pretrained on ImageNet-1k dataset [43], as the encoder network for feature extraction. Xception65 can be generally divided into 4 layers according to the size of feature maps, as illustrated in Figure 2. Let *output_stride*: OS [13, 15] denote the ratio between the size of input image and the final encoder output resolution. OS = 16 is adopted by applying atrous convolutions [13, 15] with dilation rate of 2 in Layer-4.

## 4. Experiments and Results

We evaluate the proposed architecture on two public datasets: PASCAL VOC 2012 [20] and Cityscapes [17]. The performance is measured in terms of pixel mean Intersection-over-Union (mIoU). We firstly introduce the datasets and implementation details. Thereafter we investigate the contribution of each proposed component. Finally, the comparisons with start-of-the-art approaches are presented.

### 4.1. Datasets and Implementation Details

**PASCAL VOC 2012:** The PASCAL VOC 2012 dataset consists of 20 foreground object classes and one background class. The original dataset includes 1,464 (*train*),

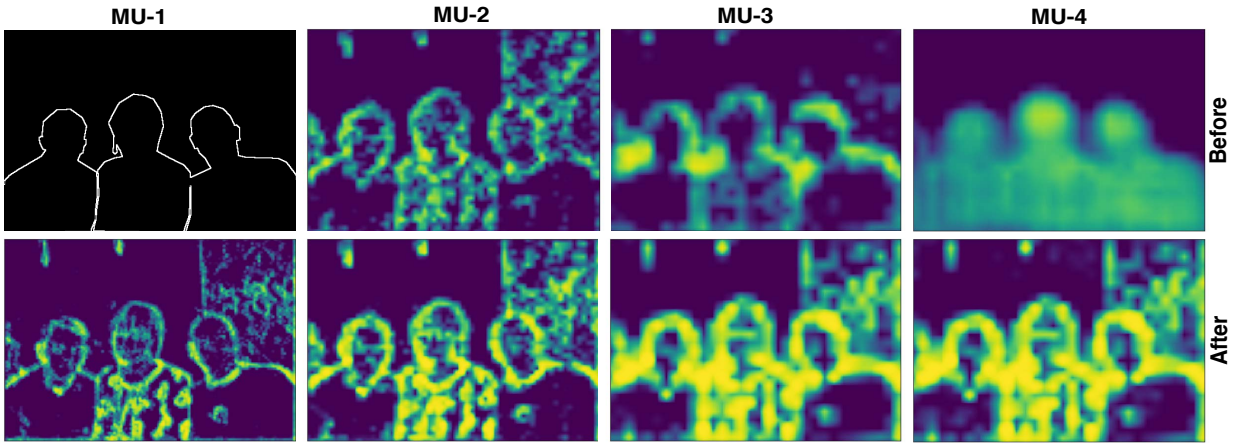



Figure 5: Intermediate contour predictions before (top row) and after (bottom row) CGCE. CGCE enhances the boundary awareness of the deeper layer features, especially the boundary attention of MU-4.

1,449 (*val*) and 1,456 (*test*) images with pixel-level annotations. The dataset is augmented by [25], contributing 10,582 (*trainaug*) training images.

**Cityscapes:** The Cityscapes dataset is a large, diverse set of high resolution 2048×1024 streets scene images from 50 different cities. The dataset contains 30 classes, and 19 of them are considered to train and evaluate our method. Cityscapes consists of 5,000 images with high quality pixel-level annotations, and 19,998 additional images with coarse annotations.

**Implementation Protocol:** Our implementation is built on TensorFlow [23]. We employ a "poly" learning rate policy where the learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with power 0.9 and initial learning rate $4e^{-3}$. A gradient multiplier of 10 is applied to the gradient of decoder to accelerate the training procedure. The network with mini-batch stochastic gradient descent (SGD), momentum 0.9, weight decay $4e^{-5}$. Dropout = 0.9 operators are applied in this work to accelerate training and avoid over-fitting. Moreover, we adopt dataset augmentation by random scaling on 6 scales $\{0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$ and random horizontal flipping during training for all both datasets.

## 4.2. Ablation Study

In this section, we investigate the contribution of each component introduced in our architecture through ablation study.

### Baselines

As a naive decoder design, adding a bilinear upsampling layer with a factor of 16 is considered as one of the baseline network, *i.e*. *Xception65-BU*, which attains the performance of 71.07% on PASCAL VOC 2012 *val* set. We adopt DeepLabv3+ [15] as the second baseline network which has performance of 79.93% on PASCAL VOC 2012 *val* set.

| Method | mIoU (%) |
|---|---|
| Xception65-BU | 71.07 |
| DeepLabv3+ | 79.93 |
| Xception65-SB | 78.61 |
| Xception65-SB-CB | 79.16 |
| Xception65-SB-CB-CGCA | 80.58 |
| Xception65-SB-CB-CGCA-CGCE | 80.86 |

Table 1: Ablation studies of our proposed architecture on PASCAL VOC 2012 *val* set. Xception65-BU and DeepLabv3+ are two baseline networks in this paper. BU: Bilinear Upsampling, SB: Semantic Branch, CB: Contour Branch, CGCA: Cross-Granularity Categorical Attention, and CGCE: Cross-Granularity Contour Enhancement.

### Semantic Branch

We firstly investigate the impact of introducing the basic mesh units SB without advanced features of CB, CA, CGCA and CGCE on top of our baseline *Xception65-BU*. We observe a performance improvement from 71.07% to 78.61% when the basic mesh unit with SB is chosen, as shown in Table 1. We owe this performance gain to the unit-wise category-orientated feature learning. Despite of the lack of categorical information propagation mechanism between the mesh units, each mesh unit adapts the feature map from encoder to focus on category related subspace which in turn improves the final semantic prediction.

### Contour Branch and Contour Attention:

Boundary information is proven effective for localizing objects in space and scale which in turn provide better boundary delineation and shape context cues for semantic segmentation task against within-class variations [51, 53, 55]. Contour branch (CB) improves performance from 78.61% to 79.16% by complementing the SB with the element-wise

contour attention (CA) unit. This performance gain agrees with our hypothesis that embedding contour information provides better boundary delineation and shape context cues for semantic segmentation.

**Cross-Granularity Categorical Attention:**

In order to enforce consistent semantic knowledge across the feature hierarchy to explicitly address the semantic gap issue, CGCA unit is added to the aforementioned architecture. Its efficacy can be clearly observed from Table 1, where the performance is improved from 79.16% to 80.58%. This quantitative gain coincides with the visual interpretation of CGCA in Figure 3, where CGCA successfully mitigates the erroneous semantic information present in features by propagating a top-down category-orientated information flow in the form of attentions.

**Cross-Granularity Contour Enhancement:**

Cross-granularity rich contour cues from shallow mesh units are propagated to compensate the poor boundary delineation at deeper layers. It significantly enhances the boundary awareness of feature representations, especially at the deeper layers *e.g*. MU-4 in Figure 5. As shown in Table 1, the contour enhancement by CGCE boosts the segmentation accuracy from 80.58% to 80.86%.

As suggested in [18], mIoU might be a good measure for region-based accuracy whereas its value overlooks how well the segmentation algorithm is at delineating object boundaries, which, nonetheless is one of the most crucial aspects of segmentation accuracy. We argue that the actual segmentation quality improvement of adding our proposed CGCE is higher than that is measured by mIoU score. To reflect this improvement, we measure the segmentation accuracy along the object boundaries with the trimap experiments in Section 4.3. Additionally, our qualitative improvement can be observed in the supplementary material.

## 4.3. Performance Evaluation along Object Boundaries with Trimap Bands

In this section, we extensively evaluate the segmentation accuracy of each architecture component from ablation study along the object boundaries with the trimap experiments [30, 11, 15]. Specifically, we apply distance transform [21] operations on the "void" label annotations which generally appear along the *val* set objects boundaries. We noticed that not all the "void" labels in the *val* set concur with object boundaries. As shown in Figure 6b, some "void" area are annotated to block objects for not being considered during evaluation. In order to quantify the accuracy of discussed methods near object boundaries more accurate, we further execute the distance transform operations on the sobel edges (Figure 6c) of *val* dataset. The distance transformation generates a staircase distance map which extends out pixel by pixel from the center object boundaries. We compute the mIoU within the trimap bands of center

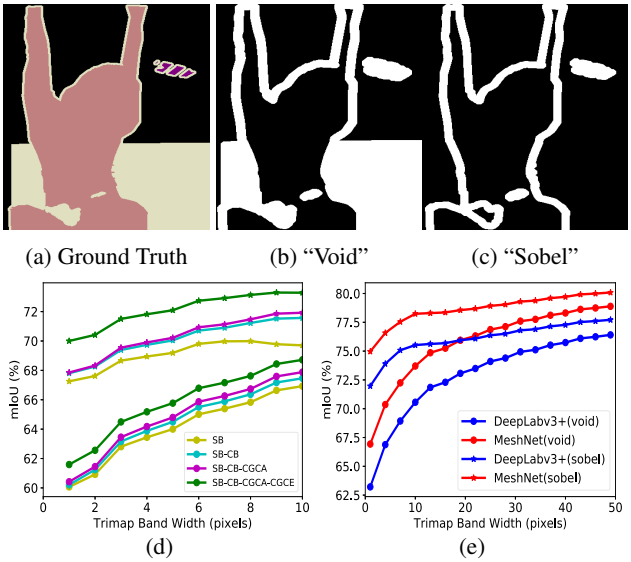



Figure 6: (a) Ground truth. (b) "Void" 20-pixel trimap. (c) "Sobel" 20-pixel trimap. (d) mIoU accuracy of architecture components from ablation studies with varying width of trimap band. The encoder "Xception65" is omitted here, and "void" boundary is marked as "$\circ$", "sobel" boundary is marked as "$\star$". (e) mIoU accuracy (with MS-COCO pre-training) of our MeshNet and DeepLabv3+ with varying width of trimap band.

| Method | train set | MS_Flip | mIoU (%) |
|---|---|---|---|
| MeshNet | | | 80.86 |
| MeshNet | ✓ | | 81.82 |
| MeshNet | ✓ | ✓ | 82.93 |

Table 2: Performance on PASCAL VOC 2012 *val* set (without MS-COCO pre-training). **MS_Flip**: Multi-Scale and left-right flipping.

"void" and "sobel" boundaries. To better visualize the improvements in close proximity to object boundaries of each discussed architecture components in Section 4.2, we calculate the mIoU within 10-pixel trimap band in Figure 6d. The top 4 curves in Figure 6d are from "sobel" boundaries and the bottom 4 curves are from "void" boundaries. As shown in Figure 6d, the method with CGCE, compared with "SB-CB-CGCA", achieves significant performance gains of 1.17% and 2.15% near "void" and "sobel" boundaries respectively on 1-pixel trimap. We also compare our best model with DeepLabv3+ [15] on 50-pixel trimap, as shown in Figure 6e. The improvement is more significant when evaluating on the narrow trimap bands, which confirms our superior quality of boundary delineations.

## 4.4. Performance Evaluation

**Performance Evaluation on PASCAL VOC 2012 Datasets**

In evaluation, we adopt multi-scaling input with scales $\{0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$ along with horizontal

| Method | mIoU (%) |
|---|---|
| FCN-8s [38] | 62.2 |
| ParseNet [36] | 69.8 |
| DeepLabv2-CRF [12] | 71.6 |
| DeconvNet [41] | 72.5 |
| DPN [37] | 74.1 |
| Piecewise [34] | 75.3 |
| LRR-CRF [22] | 75.9 |
| PSPNet [65] | 82.6 |
| DFN [62] | 82.7 |
| EncNet [63] | 82.9 |
| Res101-MeshNet | 83.5 |
| Xception65-MeshNet | **84.5** |
| MS-COCO pre-training | |
| DLC [32] | 82.7 |
| DUC [49] | 83.1 |
| RefineNet [33] | 84.2 |
| ResNet-38 [56] | 84.9 |
| PSPNet [65] | 85.4 |
| DeepLabv3(OS=8) [13] | 85.7 |
| EncNet [63] | 85.9 |
| DFN [62] | 86.2 |
| DIS [39] | 86.8 |
| DeepLabv3+(OS=8) [15] | 87.8 |
| Xception65-MeshNet(OS=16) | **87.6** |

Table 3: Performance on PASCAL VOC 2012 *test* set.

flipping operation. As PASCAL VOC 2012 dataset provides *train* set with higher quality annotations than the augmented dataset provided by [25], our architecture is further fine-tuned on *train* set before the evaluation on *val* set. The quantitative and qualitative results are shown in Table 2 and the supplementary material respectively.

For evaluation on *test* set, we further fine-tune MeshNet on PASCAL VOC 2012 *trainval* set. As a result, our proposed method achieves performance of 84.5% and 87.6% on PASCAL VOC 2012 *test* set without and with pre-training on additional MS-COCO dataset [35]. We compare with state-of-the-art methods on PASCAL VOC 2012 *test* set, and the results are listed in Table 3. It is worth noting that all our MeshNet methods are trained with *output_stride*: OS=16 while DeepLabv3 and DeepLabv3+ has OS=8 during training.

### Performance Evaluation on Cityscapes Datasets

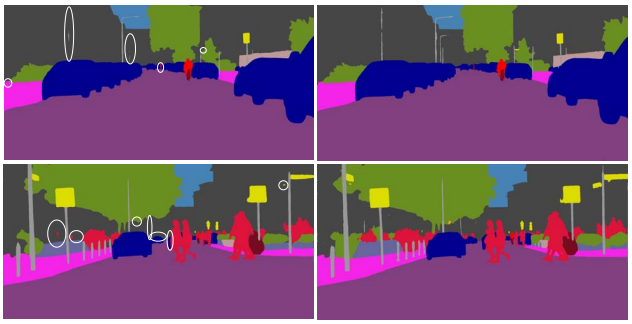
We also evaluate our architecture on Cityscapes dataset. In training, the crop size is 769×769. The quantitative and qualitative results are presented in Table 4, Figure 7 and the supplementary material respectively in comparison with state-of-the-art methods.

## 5. Conclusions

We have proposed probably one of the first deep end-to-end trainable semantic segmentation architectures with

| Method | mIoU (%) |
|---|---|
| FCN-8s [38] | 65.3 |
| DPN [37] | 66.8 |
| DeepLabv2-CRF [12] | 70.4 |
| Piecewise [34] | 71.6 |
| RefineNet [33] | 73.6 |
| DUC [49] | 77.6 |
| PSPNet [65] | 78.4 |
| BiSeNet [61] | 78.9 |
| DFN [62] | 79.3 |
| DenseASPP [60] | 80.6 |
| Res101-MeshNet | 79.4 |
| Xception65-MeshNet | **80.7** |

Table 4: Performance on Cityscapes test set.



(a) DenseASPP (b) MeshNet

Figure 7: Example results on Cityscapes dataset. Ellipses highlight the fine structures and small objects which are mis-segmented by DenseASPP, whereas our MeshNet produces regions with strong semantic coherence and accurate boundary delineation.

focuses on bridging the semantic gap and promoting boundary awareness in a unified framework. These have been long standing problems in semantic segmentation and yet are largely ignored by the state-of-the-art architectures. To this end, we have proposed a categorical attention mechanism leveraging the deep supervisions to impose semantic consistency across multi-granularity feature hierarchy. We further explicitly integrated a contour detection branch in our architecture to enhance the boundary awareness of the semantic feature in the form of element-wise contour attention at each feature hierarchy. Additionally, we introduce a cross-granularity contour enhancement mechanism to propagate rich boundary cues from early layers to deep layers. These novel contributions delivered significantly improved region-wise semantic coherency and accurate object contour localization. We have performed extensive evaluations of our architectures and obtained state-of-the-art performance on challenging datasets.

## References


[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389. IEEE, 2015.

[5] I. Biederman and G. Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988.

[6] P. Bilinski and V. Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6596–6605, 2018.

[7] H. Caesar, J. Uijlings, and V. Ferrari. Cocostuff: Thing and stuff classes in context. *CoRR, abs/1612.03716*, 5:8, 2016.

[8] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2488–2496, 2015.

[9] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[10] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[14] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[16] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1251–1258, 2017.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[18] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation?. In *Proceedings of the British Machine Vision Conference*, volume 27, page 2013. Citeseer, 2013.

[19] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[21] R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):2, 2008.

[22] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, pages 519–534. Springer, 2016.

[23] S. S. Girija. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.

[24] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[25] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. 2011.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[27] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7132–7141, 2018.

[28] J.-J. Hwang and T.-L. Liu. Pixel-wise deep learning for contour detection. *arXiv preprint arXiv:1504.01989*, 2015.

[29] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding long-short term memory for image caption generation. *arXiv preprint arXiv:1509.04942*, 2015.

[30] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[31] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[32] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3202, 2017.

[33] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation.

[34] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[36] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[37] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[39] P. Luo, G. Wang, L. Lin, and X. Wang. Deep dual learning for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2718–2726, 2017.

[40] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

[41] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[42] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[44] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[45] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2015.

[46] H. Wang and T. Wang. Primary object discovery and segmentation in videos via graph-based transductive inference. *Computer Vision and Image Understanding*, 143:159–172, 2016.

[47] H. Wang, T. Wang, K. Chen, and J.-K. Kämäräinen. Cross-granularity graph inference for semantic video object segmentation. In *IJCAI*, pages 4544–4550, 2017.

[48] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik. Training deeper convolutional networks with deep supervision. *arXiv preprint arXiv:1505.02496*, 2015.

[49] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 1451–1460. IEEE, 2018.

[50] T. Wang, J. Collomosse, D. Slatter, P. Cheatle, and D. Greig. Video stylization for digital ambient displays of home movies. In *Proceedings of NPAR*, pages 137–146. ACM, 2010.

[51] T. Wang, B. Han, and J. Collomosse. Touchcut: Single-touch object segmentation driven by level set methods. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 881–884. IEEE, 2012.

[52] T. Wang, B. Han, and J. P. Collomosse. Touchcut: Fast image and video segmentation using single-touch interaction. *Computer Vision and Image Understanding*, 120:14–30, 2014.

[53] T. Wang and H. Wang. Graph transduction learning of object proposals for video object segmentation. In *Proceedings of Asian Conference on Computer Vision*, pages 553–568. Springer, 2014.

[54] T. Wang, H. Wang, and L. Fan. Robust interactive image segmentation with weak supervision for mobile touch screen devices. In *Proceedings of International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2015.

[55] T. Wang, H. Wang, and L. Fan. A weakly supervised geodesic level set framework for interactive image segmentation. *Neurocomputing*, 168:55–64, 2015.

[56] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[57] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[58] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1395–1403, 2015.

[59] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[60] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes.

[61] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *arXiv preprint arXiv:1808.00897*, 2018.

[62] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. *arXiv preprint arXiv:1804.09337*, 2018.

[63] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[64] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[66] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4. IEEE, 2017.