Context and Attribute Grounded Dense Captioning

Guojun Yin[1], Lu Sheng[2,4], Bin Liu[1]*, Nenghai Yu[1], Xiaogang Wang[2], Jing Shao[3]
[1]University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information,
The Chinese Academy of Sciences, [2]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong,
[3]SenseTime Research, [4]College of Software, Beihang University
gjyin@mail.ustc.edu.cn, lsheng@buaa.edu.cn,
{flowice,ynh}@ustc.edu.cn, xgwang@ee.cuhk.edu.hk, shaojing@sensetime.com

Abstract

Dense captioning aims at simultaneously localizing semantic regions and describing these regions-of-interest (ROIs) with short phrases or sentences in natural language. Previous studies have shown remarkable progresses, but they are often vulnerable to the aperture *problem that a caption generated by the features inside one ROI lacks* contextual coherence *with its surrounding context in the input image. In this work, we investigate contextual reasoning based on multi-scale message propagations from the neighboring contents to the target ROIs. To this end, we design a novel end-to-end context and attribute grounded dense captioning framework consisting of 1) a contextual visual mining module and 2) a multi-level attribute grounded description generation module. Knowing that captions often co-occur with the linguistic attributes (such as* who, what *and* where*), we also incorporate an auxiliary supervision from hierarchical linguistic attributes to augment the distinctiveness of the learned captions. Extensive experiments and ablation studies on Visual Genome dataset demonstrate the superiority of the proposed model in comparison to the state-of-the-art methods.*

1. Introduction

Dense captioning, which was first introduced by [20], is to parse semantic contents in an input image and describe them with captions in natural languages. It can benefit other tasks, including image captioning [38], segmentation [28], visual question answering [14] and *etc*. In this paper, we mainly focus on the caption generation and adopt Faster RCNN [29] for semantic instances localization, following recent advances [20, 34].

Differing from subjective image descriptions for high-level abstraction of an entire image, captions of semantic

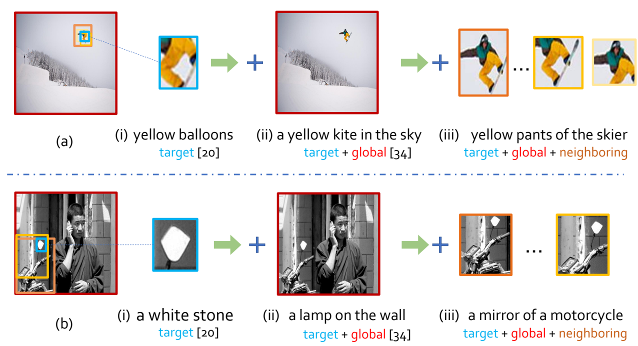

Figure 1. Dense captioning with different levels of contextual interactions: (i) without any contextual cues (marked by blue) [20], (ii) with guidance from the global cue (marked by red) [34], and (iii) with mutual interactions from neighboring (marked by orange) and global visual information. (Best viewed in color.)

instances in compact bounding boxes are far more objective and less affected by ambiguities raised by subjective annotations. That is, incorrect captions may be generated when the target regions are visually ambiguous without contextual reasoning. For example, it may falsely caption the target ROIs marked in blue-box as "yellow balloons" rather than "yellow pants" in Fig. 1(a-i), if not aware of their contextual visual contents [20]. An alternative solution proposed in [34] try to exploit the global feature from the entire image as the contextual cue to improve the region captioning. However, the descriptions sometimes are corrupted by global appearance, especially for small and unusual objects against dominant global contents. The "yellow pants" in Fig. 1 (a-ii) is mistakenly described as "yellow kite in the sky". The similar phenomenon happens in Fig. 1 (a-ii) that it mistakenly generates "a mirror" rather than "a lamp".

In contrast to the prior arts, in this study, we show that the innovative model, named as *Context and Attribute Grounded Network* (CAG-Net), designed with contextual correlated visual cues (*i.e*., *local, neighboring, global*) permits multi-scale contextual message passing to reinforce regional description generation. For example, the neighbor-

*Bin Liu is the corresponding author.

ing ROIs marked in warm-box in Fig. 1(a-iii), semantically connecting to the visual features in the target in blue-box in Fig. 1(a-i), provide more valuable hints that the target is a "yellow pants" belonging to a skier. Such contextual learning has shown its remarkable potential in other tasks including object detection, segmentation and retrieval. However, the learning of contextual representation, and how it can effectively function on dense captioning, remains an open problem. Specifically, the proposed CAG-Net consists of two vital modules:

1) *Contextual Feature Extractor*, establishing a non-local similarity graph for the feature interaction between the target ROI and its neighboring ROIs based on their feature affinity and spatial nearness, allows adaptive contextual information sharing from multiple adjacent ROIs (*i.e*., global and neighbors) to interact with the target ROI.

2) *Attribute-Grounded Caption Generator* adopts LSTM as the fundamental unit and fuses contextual features to generate the description for the target ROI. To reinforce the coarse-to-fine structure of description generation, we adopt coarse-level and fined-level linguistic attribute losses as the additional supervision respectively at the sequential LSTM cells. Without sequential restrictions from the ground-truth captions, such keywords or attributes are more recognizable by the content in the target ROI, and thus own a more stable discriminative power for the extraction of visual patterns. To some extent, it is similar with the visual attributes of objects in multi-label classification.

Our contributions are listed as follows:

1) We design a context and attribute grounded dense captioning model that permits multi-scale (*i.e*., local, neighboring, global) contextual information sharing and message passing, where the knowledge integration is built on a non-local similarity graph among instances in the input image.

2) A coarse-to-fine linguistic attribute supervision is proposed to enhance the discriminativeness of the generated captions, in which the ground-truth hierarchical linguistic attributes are matched to the predicted keywords through a novel coarse-to-fine manner.

3) Extensive experiments demonstrate the effectiveness of the proposed CAG-Net model on the challenging large-scale VG dataset.

2. Related Work

Image captioning to describe a general image with natural language was explored in recent years [5, 26, 12, 30, 2, 27, 35, 25]. The works [5, 36, 1, 6, 19, 7] focused on improve ifiguremage captioning by the attention-embedded features generated by an additional attention model. Based on the attention model, Gu *et al*. [15] adopted a coarse-to-fine framework which increased the model complexity gradually with increasingly refined attention weights for image captioning. In our work, dense image captioning renders individual captions for different ROIs in the image. As for dense captioning, we firstly adopt the multi-scale feature interaction and attribute grounded generation for accurate region descriptions. Our coarse-to-fine strategy is based on the hierarchical attribute supervisions rather than the different attention inputs of the description generation modules [15]. The previous works [38, 36] adopted the attributes (the words in the vocabulary) to train an additional model for another input of the LSTM cells for the description generation. Differing from that, our work adopts the linguistic attributes as the auxiliary supervision for coarse-to-fine generation without any external branches or input.

Dense Image Captioning. Dense image captioning is supposed to not only localize the regions of interest in the image but also generate descriptions with natural language, which was first proposed in [20]. Johnson *et al*. [20] introduced a new dense localization layer, which was fully differentiable and used bilinear interpolation to smoothly extract the activations inside each region. Yang *et al*. [34] exploited more accurate localization for regions by joint inference of localization and description for a given region proposal, while the global feature of the image was used as the contextual cues to improve region captioning. However, these previous works did not capture the relative features of different regions and valid message passing between contextual regions for accurate region captioning.

Contextual Learning. Contextual learning was employed in various topics in recent years [32, 24, 22, 39, 33, 10, 9], *e.g*., object detection, segmentation, and retrieval. For both detection and segmentation, learning feature representations from a global view rather than the located object itself has been proven effective by [37, 28]. Gkioxari *et al*. [13] used more than one region proposals for action recognition while Hu *et al*. [17] processed a set of objects simultaneously through interaction between their appearance feature and geometry, thus allowing modeling of their relations for object detection. As for contextual learning for image captioning, Yao *et al*. [35] computed the probability distribution on all the semantic relation classes for each object pair with the learnt visual relation classifier to establish the semantic graph for image captioning. Contextual feature learning among the located regions for dense captioning has never been explored in the previous works. In our work, we establish contextual message passing module without additional branches or any auxiliary relation labels.

3. Context and Attribute Grounded Dense Captioning (CAG-Net)

In this paper, we propose a novel end-to-end dense image captioning framework, named as *Context and Attribute Grounded Dense Captioning* (CAG-Net). As shown in

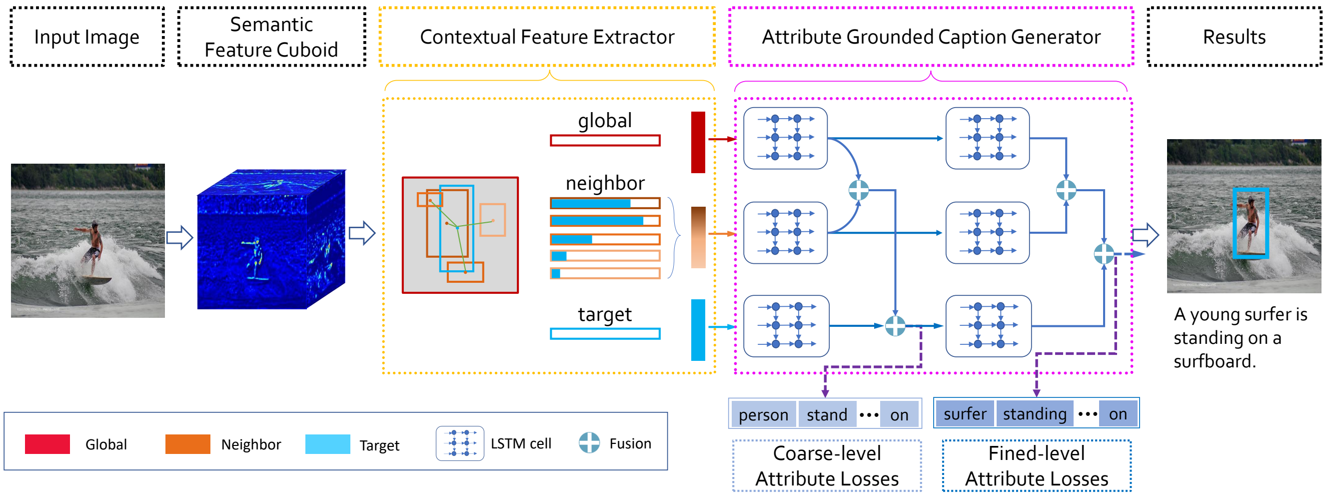

Figure 2. The architecture of CAG-Net. The multi-scale features are generated by the proposed Contextual Feature Extractor after region proposals. Then the *local* (in blue) feature of the target region and multi-scale context cues, *i.e.*, *global* (in red) and *neighboring* (in orange), broadcast into the Attribute Grounded Caption Generator for region captioning in parallel. The final descriptions of the target region are generated jointly by the hierarchical structures trained with the auxiliary attribute losses.

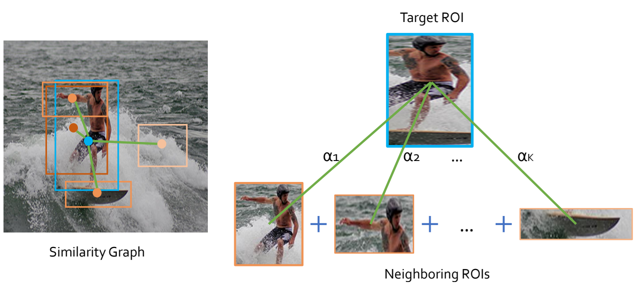

Figure 3. An example of Contextual Feature Extractor for the target proposal. (left) The similarity graph between target (in blue) proposal and contextual neighboring (in orange) proposals are generated considering both spatial configuration and appearance similarity. (right) The *neighboring* feature are obtained by fusing the contextual neighboring proposals with the similarity graph. Best viewed in color.

Fig. 2, we first learn visual features of the input image by a CNN model as the way adopted by Faster RCNN [29], and obtain the semantic features. Such semantic features are used to generate a set of candidate regions (ROIs) by a Region Proposal Network (RPN) [29]. Based on these ROI features, we introduce a *Contextual Feature Extractor* (CFE) which generates the global, neighboring and local (*i.e.*, target itself) cues (Sec. 3.1). The neighboring cues are collected by establishing a similarity graph between the target ROI and the neighboring ROIs, shown in Fig. 3. The multi-scale contextual cues, broadcast in parallel, are fused by an *Attribute Grounded Caption Generator* (AGCG) which employs multiple LSTM [16] cells (Sec. 3.2). To generate rich and fine-grained descriptions and reinforce the coarse-to-fine procedure of description generation, we adopt an auxiliary supervision, *Linguistic Attributes*, hierarchically upon the outputs of the sequential LSTM cells, as in Fig. 2. The proposed model is trained to both minimize the sentence loss and the binary cross-entropy losses (attribute losses) for caption generation.

3.1. Contextual Feature Extractor

Denote the regions-of-interest (ROIs) in an image as $\mathcal{R} = \{\mathbf{R}_i | i = 1, 2, ..., N\}$ and the entire image as $\mathbf{R}^*$. The contextual features for the local region $\mathbf{R}_i$ are from the multi-scale contextual cues of *local* region $\mathbf{R}_i$, *neighboring* region $\mathcal{R}_i^n = \mathcal{R}/\mathbf{R}_i$, and the *global* region $\mathbf{R}^*$. For the target region $\mathbf{R}_i$, denote the *local, neighboring* and *global* features as $\mathbf{F}_i^l$, $\mathbf{F}_i^n$, and $\mathbf{F}_i^g$, respectively, where $\mathbf{F}_i^g$ refers to the features extracted from the entire input image and $\mathbf{F}_i^l$ is the feature of the target instance. The *Contextual Feature Extractor* (CFE) focuses on exploring the neighboring features $\mathbf{F}_i^n$ which can be formulated as $\mathbf{F}_i^n = f(\mathbf{R}_i, \mathcal{R}_i^n)$.

We design a region-level similarity graph (*i.e.*, ROI-level) for neighboring ROIs aggregation, inspired by pixel-level non-local operations. Non-local means [4] has been often used as a filter by computing a weighted mean of all pixels in an image, which allows pixels to contribute to the filtered response based on the patch appearance similarity. Similarly, neighboring ROIs with similar semantic appearance are supposed to contribute more on the feature extraction for the target local instance. Following the operation in [4], we rewrite the formulation of $f(\mathbf{R}_i, \mathcal{R}_i^n)$ as

$$f(\mathbf{R}_i, \mathcal{R}_i^n) = \sum_{\forall j, j \neq i} \mathcal{G}(\mathbf{F}_i^l, \mathbf{F}_j^l)\mathbf{F}_j^l, \tag{1}$$

where $\mathcal{G}(\mathbf{F}_i^l, \mathbf{F}_j^l)$ is the appearance similarity between region $\mathbf{R}_i$ and $\mathbf{R}_j$, and $\mathbf{F}_i^l$ is the fixed-length local feature of

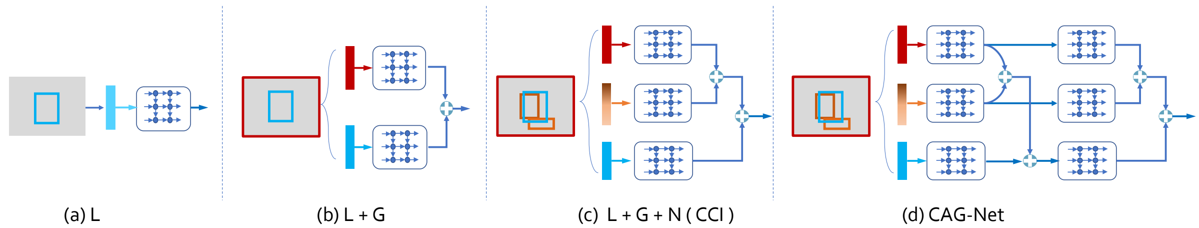

Figure 4. Comparisons between different network structures. (a) L generates the descriptions separately after region proposals; (b) L + G generates descriptions with not only the local feature but also the global feature of the image; (c) L + G + N (CCI) integrates global, neighboring and local information for the target to generate descriptions; (d) CAG-Net by multiple LSTM cells is a stacked version of (c) CCI but supervised with hierarchical linguistic attribute losses.

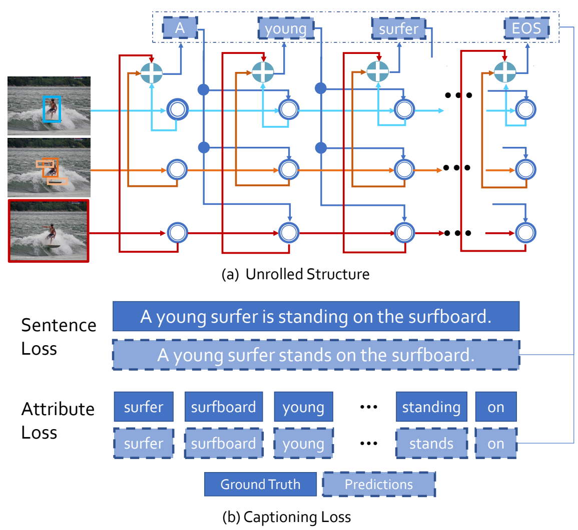

Figure 5. The unrolled structure of Contextual Cue Integrator (CCI). (a) Unrolled structure integrates the local (in blue) information and multi-scale context cues, *i.e.*, global (in red) and neighboring (in orange). The hollow circle stands for the LSTM cell while the plus sign for the feature fusion briefly. (b) The captioning loss consists of a sentence loss and an attribute loss.

region $\mathbf{R}_i$. The similarity $\mathcal{G}$ is the normalized cross correlation based on Gaussian function, formulated as,

$$\mathcal{G}(\mathbf{F}_i^l, \mathbf{F}_j^l) = \frac{exp({\mathbf{F}_i^l}^\top \mathbf{F}_j^l)}{\sum_{\forall j, j \neq i} exp({\mathbf{F}_i^l}^\top \mathbf{F}_j^l)}, \quad (2)$$

where ${\mathbf{F}_i^l}^\top \mathbf{F}_j^l$ is dot-product similarity of cross correlation. Therefore, we can obtain the similarity graph for each target ROI with its neighboring ROIs in the image.

General object detection algorithm usually generates redundant region candidates (ROIs) to ensure the accuracy and robustness in region localization and detection. However, in this case, the integrated *neighboring* feature $\mathbf{F}_i^n$ will be contaminated by distant and independent proposals, and the amount of ROIs in $\mathcal{R}_i^n$ also tremendously increase the computation cost and noises in the environment. Therefore, we sample a subset of $\mathcal{R}_i^n$ based on their spatial nearness such that the closer ROIs are more relative to the target ROI. We sort ROIs in $\mathcal{R}_i^n$ based on the IoU (intersection over union) metric with the target region $\mathbf{R}_i$. By sampling the top-k proposals as $\hat{\mathcal{R}}_i^n$, the calculation of the *neighboring* features can be accelerated as $\mathbf{F}_i^n = f(\mathbf{R}_i, \hat{\mathcal{R}}_i^n)$.

3.2. Attribute Grounded Caption Generator

We present a novel caption generator with two parts: (1) a *Contextual Cue Integrator* to fuse contextual features produced by the CFE in Sec. 3.1, and (2) an *Attribute Grounded Coarse-to-Fine Generator* with coarse-level and fined-level linguistic attribute losses as the additional supervision to enhance the discriminativeness of the generated captions.

Contextual Cue Integrator (CCI) - The contextual cue integrator adopts multiple LSTM cells to hierarchically integrate the multi-scale contextual features into the localized features. The *local, neighboring* and *global* features are spread through the LSTM cells so as to generate context-aware descriptions for the target ROI. These descriptions are fused together for the final captioning of the target region at each time step of LSTM. The unrolled contextual cue integration module is shown in Fig. 5(a). The *local* branch is regarded as the backbone for the target and the *global* and *neighboring* branches are grouped as multi-scale contextual cues to provide complementary guidances. Thus, the contextual cues are adaptively combined at first, and they are then added to the local branch via a second adaptive fusion, as shown in Fig. 4(c). The importance of different levels' features is regularized by the adaptive weights, which are optimized during training the framework.

Attribute Grounded Coarse-to-Fine Generator - It is challenging in generating rich and accurate descriptions just by the sequential LSTMs. To this end, we increase its representative power by introducing a coarse-to-fine caption generation procedure with sequential LSTM cells, *i.e.*, coarse stage and refined stage supervised with the auxiliary hierarchical linguistic attribute losses.

The linguistic attribute losses serve as the intermediate and auxiliary supervisions from coarse to fine in addition to the general sentence loss of captioning, implemented at

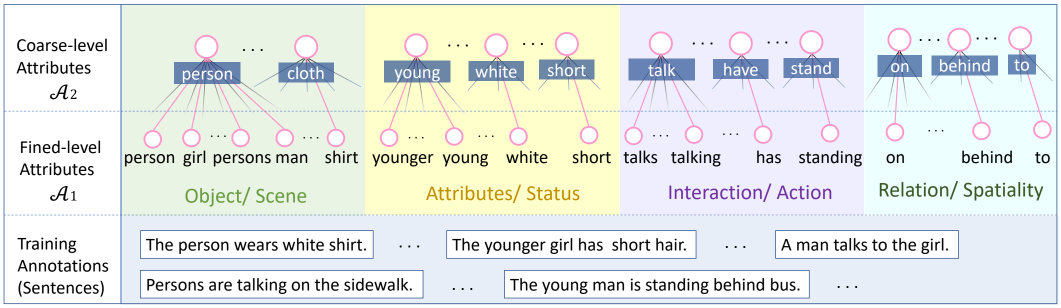

Figure 6. Illustration of sentence itemization. Fined-level attributes $\mathcal{A}_1$: the original sentences of training annotations (bottom) are itemized to individual words and divided into four groups: object/scene (noun), attribute/status (adjective), interaction/action (verb) and relation/spatially (preposition). Coarse-level attributes $\mathcal{A}_2$: the individual words are normalized and clustered by semantical similarity for high-level words, *e.g.*, the *girl* and *man* in $\mathcal{A}_1$ belong to *person* in $\mathcal{A}_2$.

each stage as shown in Fig. 2. The attribute losses are formatted as binary classification (*i.e.*, exist or not) losses for each attribute separately during the training procedure. As shown in Fig. 5 (b), the attributes, *e.g.*, *surfer, standing, young* and *on*, will be measured individually regardless of the speech order, similar as the multi-label classification for attribute recognition of objects.

The subsequent LSTM layer (`refined` stage) is supposed to serve as the fine-grained decoders for the coarse regional descriptions generated by the preceding one (`coarse` stage). The hidden vectors of LSTM cells produced by the coarse stage are taken as the disambiguating cues to the refined stage. The outputs of *global* and *neighboring* branches at the coarse stage are used as the inputs of the respective branches directly at the refined stage. The adaptive fusion of these three branches at the coarse stage is fed as the input at the refined stage. Meanwhile, these vectors are used for coarse-level attribute prediction. The connection of the branches at the multiple stages is shown in the Fig. 4(d). The final outputs of the word decoder at the refined stage are the generated descriptions for the target region. Meanwhile, these outputs are used for the fined-level attribute prediction as well.

These linguistic attributes are predicted from the outputs of the LSTMs during the training procedure and the unsolved problem here is how to get the ground-truth linguistic attributes. In our work, the hierarchical *linguistic attributes* are obtained by itemizing the sentences in the training split with natural language processing toolkit (NLTK).

1) Fined-level attributes $\mathcal{A}_1$ for `refined` stage. We distill the linguistic knowledge from the training annotations (sentences or phrases) to individual keywords/attributes, by the speech toolkit from NLTK , as shown in Fig. 6. The reference sentences are parsed into four groups by the part-of-speech, *i.e.*, nouns, adjectives, verbs and prepositions from the following aspects respectively: (1) The noun words are usually the labels of objects or scenes, *e.g.*, *person*, *bus*, *sidewalk* and *etc.*; (2) adjectives represent attributes or status, *e.g.*, *young*, *black*; (3) verbs are meanings of actions or interactions, *e.g.*, *standing*, *talks*; (4) prepositions for relations or spatiality, *e.g.*, *behind*. The fined-level attributes like *surfer* and *standing* are used at the latter stage for the exact information extraction.

2) Coarse-level attributes $\mathcal{A}_2$ for `coarse` stage. We use the high-level semantically clustered attributes, *e.g.*, *person*, *stand* to stand for the major information. We observe that labels with the same concept may have different singular and plural forms or different participles, *e.g.*, *persons* versus *person*, *talks* versus *talking*. These words are normalized to a unified format by NLTK Lemmatizer, *e.g.*, *talk* from *talks* and *talking*. Furthermore, labels having closer semantic correlation (*e.g.*, *girl* and *man* are hyponyms of *person*) need to be distinguished from other semantic concepts like *cloth*, as shown in the top panel of Fig. 6. Therefore, we cluster the labels with their semantical similarities computed by Leacock-Chodorow distance [31]. We find a threshold of 0.85 is well-suited for splitting semantic concepts. The coarse-level items like *person* and *stand* are used at the preceding stage for the key information extraction.

4. Experiments

4.1. Experiment Settings

Dataset. Visual Genome (VG) region captioning dataset [21] is used as the evaluation benchmark in our experiments. For fair comparisons, we use the dataset of version 1.0 and the same train/validation/test splits as in [20], *i.e.*, 77398 images for training and 5000 images each assigned for validation and test.

Evaluation Metric. Following [20], the mean Average Precision (mAP) are measured across a range of thresholds for both accurate localization and language description, inspired by the evaluation metrics in object detection [11, 23] and image captioning [3]. For localization, intersection over union (IoU) thresholds .3, .4, .5, .6, .7 are used while METEOR [3] score thresholds 0, .05, .1, .15, .2, .25 used for language similarity. The average precision is measured across all pairwise settings, *i.e.*, *(IoU, METEOR)*, of these methods and report the mean AP (mAP), which means the mAP is computed for different IoU thresholds for localization accuracy, and different METEOR score thresholds for language similarity, then averaged as the final score.

To isolate the accuracy of language in the predicted captions without localization, the predicted captions are evaluated neglecting their spatial positions. Following [20], the references of each prediction are generated by merging ground truth across each image into a bag of reference sentences. Apart from the mAP score introduced above, the METEOR score will be reported as the auxiliary evaluation metric, denoted as *METEOR*. Note that the references from all regions in an image only offer the global and coarse ground truth descriptions.

Implementation Details. We use VGG-16 [21] pretrained on ImageNet [8] as the network backbone. In Fig.2, we use 6 LSTM cells in total, *i.e.*, one LSTM for *local, neighboring, global* features respectively at each stage. The newly-introduced layers and LSTM cells are randomly initialized and our proposed CAG-Net is end-to-end trained. The implementations are based on Faster RCNN [29] using Caffe [18], and the networks are optimized via stochastic gradient descent (SGD) with base learning rate as 0.001. The input image is re-sized to have a longer side of 720 pixels and 256 proposals are sampled per image in each forward pass of training. The LSTM cell for sequential modeling has 512 hidden nodes. The most $10,000$ frequent words in the training annotations are remained as the vocabulary and other words are collapsed into a special $<$`UNK`$>$ token under the same conditions as in [34]. Following [20], we discard all sentences with more than 10 words (7% of annotations), that is the time length of the LSTMs is 10.

The losses of our framework are from two aspects: 1) Location: Smooth ℓ_1 loss for bounding box regression ($\mathcal{L}_{bbox}$) and softmax loss for binary foreground/background classifier ($\mathcal{L}_{cls}$), 2) Caption: Cross entropy loss of sentences for description generation ($\mathcal{L}_{sent}$), following [34] and binary cross entropy loss for linguistic attribute recognition ($\mathcal{L}_{attr}$) . The total loss function is $\mathcal{L} = \mathcal{L}_{sent} + \alpha\mathcal{L}_{bbox} + \beta\mathcal{L}_{cls} + \gamma\mathcal{L}_{attr}$, where $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.01$ in our experiments with the empirical values.

In evaluation, we follow the settings of [20] for fair comparisons. 300 proposals with the highest predicted confidence are remained after non-maximum suppression (NMS) with IoU threshold 0.7. We use efficient beam-1 search to produce region descriptions, where the word with the highest confidence is selected at each time step. With another round NMS with IoU threshold 0.3, the remaining regions and their generated descriptions are used for the final evaluation. To establish an upper bound regardless of region proposals, we evaluate the models on ground truth bounding boxes as well, marked as GT in the tables.

Methods	RPN		GT	
	mAP	METEOR	mAP	METEOR
CAG-Net	**10.51**	**0.279**	**36.29**	**0.316**
T-LSTM [34]	9.31	0.275	33.58	0.307
FCLN [20]	5.39	0.273	27.03	0.305

Table 1. Quantitative results on Visual Genome comparing with state-of-the-art methods, T-LSTM [34] and FLCN [20]. Results in bold indicate the best performance. The metrics on T-LSTM, *i.e.*, METEOR, are not provided in the paper and we measure these metrics using the model provided by the authors.

Methods		CAG-Net	L+G+N	L+G	L
mAP	RPN	**10.51**	9.55	7.97	6.31
	GT	**36.29**	33.50	31.77	26.70

Table 2. Ablation study on CAG-Net compared the variants of the contextual cue integration module, *i.e.*, 1) `L`, local cue without neighboring nor global features, 2) `L+G`, local and global cue integration and 3) `L+G+N`, local, global and neighboring integration without stacking contextual cue integration modules. Results in bold indicate the best performance.

4.2. Comparison with State-of-the-Art Methods

We quantitatively compare the performances of the proposed *Context and Attribute Grounding Dense Captioning* (CAG-Net) model with the previous state-of-the-arts, *i.e.*, FCLN[20] and T-LSTM [34]. FCLN [20] introduces a fully differentiable layer for dense localization. The captioning per region is generated solitarily without any message passing from the contextual features. T-LSTM [34] designs network structures that incorporate two novel parts: joint inference for accurate localization and context fusion with the global scene for accurate description regardless of the interactions among the relative regions.

The comparison experiments use the same settings as the prior arts, shown in Tab. 1. The CAG-Net significantly outperforms these methods by achieving a gain on mAP score from 9.31% to 10.51% using RPN and from 33.58% to 36.29% using the ground truth bounding boxes compared to the previous state-of-the-art, T-LSTM [34]. The performance gains are mainly from the benefits of attribute grounded coarse-to-fine description generation using the contextual feature extractor and message integration among the regions. The proposed CAG-Net presents a strong capability in capturing the correlation among relative regions and generating more accurate descriptions.

It is observed that the METEOR scores of different methods are approximate while the mAP scores have a large margin. That is because that the METEOR score for the predicted caption is calculated by using all ground truth descriptions of all the regions in the image as the references. These references are coarse and may not be accurate for a certain region. In the following ablation study (Sec. 5), we mainly focus on the comparison of mAP scores.

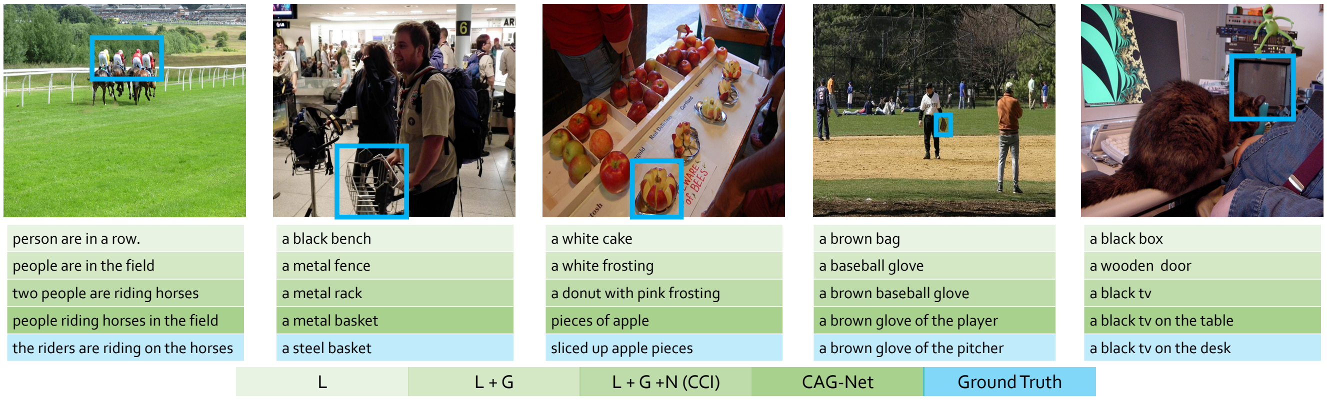

Figure 7. Qualitative results of CAG-Net compared with variants of different module configurations on VG dataset, *i.e.*, (a) L(Local Cue), (b) L+G (Local and Global Integration), (c) L+G+N (CCI) (Local, Global and Neighboring Integration).

5. Ablation Study

5.1. CAG-Net

Attribute Grounded Caption Generator with Contextual Cues. To demonstrate the benefits of multi-scale contexts and attribute grounded captioning module, we compare the results of CAG-Net in Fig. 4 (d) with the variants by removing individual cue step by step, *i.e.*, 1) `L`, local cue as the baseline without either contextual neighboring or global features as shown in Fig. 4 (a), 2) `L+G`, local and global cue integration without contextual neighboring cues in Fig. 4 (b) and 3) `L+G+N`, local, global and neighboring integration without stacking contextual cue integration modules in Fig. 4 (c), defined as CCI in Sec 3.2. The quantitative results are reported in Tab. 2.

Compare with basic `L`, the mAP of `L+G+N` can be improved from 6.31% to 9.55% using RPN and from 26.70% to 33.50% using ground truth boxes by involving contextual feature extractor and message integration while the mAP of `L+G` which only includes the global cues achieves 7.97% using RPN and 31.77% using ground truth bounding boxes. The significant improvement demonstrates the importances of contextual cue integration between multi-scale contexts and individual regions for region generation and the contextual cues, *i.e.*, *global* and *neighboring* make a certain contribution to improving the final performances. Furthermore, with the assistance of the linguistic attribute losses, the mAP of CAG-Net achieves 10.51% in mAP using RPN by a gain of 0.96% compared to `L+G+N` (CCI) while a gain of 1.79% using the ground truth bounding boxes. No doubt that the generated descriptions are more accurate and rich for the regions when adopting attribute grounded coarse-to-fine captioning module.

The qualitative results are shown in Fig. 7. The descriptions directly generated by the target regions are fallible due to lack of enough visual information, *e.g.*, mistaking the *baseball glove* for *a brown bag*, the *apple pieces* for *a white cake* and the *steel basket* for *black bench*. The involved global cues of the image also lead to deviation, *e.g.*, the *tv in the room* is mistakenly predicted as *a wooden door*, although positive effect sometimes, *e.g.*, the *glove* is accurately predicted with the assistance of the global image feature. Furthermore, the coarse-to-fine generation module will reinforce more rich descriptions *a black tv on the table* compared with individual module *a black tv* shown in the figure. The results shows the excellent performance of the proposed context and attribute grounded generation structure for dense captioning.

Methods		CAG-Net					CCI
		$(\mathcal{A}_2, \mathcal{A}_1)$	$(\mathcal{A}_1, \mathcal{A}_1)$	$(\mathcal{A}_2, \mathcal{A}_2)$	$(1k, 1k)$	$(-, -)$	
mAP	RPN	**10.51**	9.93	9.99	9.95	9.59	9.55
	GT	**36.29**	34.98	35.17	35.02	33.78	33.50

Table 3. Ablation study on CAG-Net compared the variants of linguistic attribute losses, *i.e.*, 1) $(\mathcal{A}_2, \mathcal{A}_1)$, with the proposed coarse-to-fine attributes, 2) $(\mathcal{A}_1, \mathcal{A}_1)$, only with the fined-level attributes $\mathcal{A}_1$, 3) $(\mathcal{A}_2, \mathcal{A}_2)$, only with the coarse-level attributes $\mathcal{A}_2$, 4) $(1k, 1k)$, replacing the proposed attributes with the top $1k$ attributes, 5) $(-, -)$, stacked structure without any attributes, 6) CCI, just one stage without attributes. Results in bold are the best.

Linguistic Attribute Losses. To demonstrate the benefits of the proposed linguistic attribute losses, we compare the performances of CAG-Net with the variants of linguistic attributes by 1) "$(-, -)$", removing all the auxiliary linguistic attribute losses in the framework, 2) "$(\mathcal{A}_1, \mathcal{A}_1)$", only with the fined-level attributes $\mathcal{A}_1$ at two stages, 3) "$(\mathcal{A}_2, \mathcal{A}_2)$", only with the coarse-level attributes $\mathcal{A}_2$ at two stages, 4) "($1k$,$1k$)", replacing the proposed linguistic attributes with the top $1k$ attributes (the top $1k$ most frequent words in the vocabulary) at two stages.

The results are shown in Tab. 3 and CAG-Net with the proposed coarse-to-fine linguistic attributes is denoted as "$(\mathcal{A}_2, \mathcal{A}_1)$". Compared with "CCI", CAG-Net without any attributes (denoted as "$(-, -)$" in Tab.3) gets the approximate results because the navie stacking description gen-

Methods		Random	Nearest	SG	FC	AVE	MAX
mAP	RPN	8.626	9.144	9.315	8.132	7.981	8.024
	GT	32.274	33.411	33.412	30.272	29.937	30.121

Table 4. Results of Contextual Feature Extractor with different settings. "Random" means selecting the contextual neighboring regions randomly from all the regions in the image. "Nearest" means selecting the relative regions from the nearest ones sorted by the IoU scores. "SG" means fusing these neighboring regions with similarity graph. "FC" means fusing k-sorted neighboring regions with fully connected layer. "AVE" means average-pooling of k-sorted neighboring regions. "MAX" means max-pooling of k-sorted neighboring regions.

eration modules cannot significantly improve the performance although with more parameters. In contrast, the attribute grounded structure with the proposed coarse-to-fine attributes can achieve a gain from 9.59% to 10.51%(using RPN) and from 33.78% to 36.29% (using ground truth boxes) because of the auxiliary hierarchical supervision of the proposed linguistic attribute losses. Furthermore, to evaluate the effectiveness of the coarse-to-fine structure, we compare CAG-Net, *i.e.*, "$(\mathcal{A}_2, \mathcal{A}_1)$", with the variants of different linguistic attributes, *i.e.*, '$(\mathcal{A}_1, \mathcal{A}_1)$", "$(\mathcal{A}_2, \mathcal{A}_2)$" and "($1k$, $1k$)". Without the coarse-to-fine strategy at two stages, the stacked structures with different attributes cannot achieve as good performance as CAG-Net both using RPN and using ground truth bounding boxes. It is significant that the proposed linguistic attribute losses from the coarse to fine stage can improve the description generation of target regions.

5.2. Contextual Feature Extractor

In this section, we compare the performances of Contextual Feature Extractor (CFE) with variants by changing one of the hyper-parameters or settings step by step to explore the best practice of the proposed contextual feature extractor. As for the generation structure, we use CCI instead of CAG-Net due to the faster speed and less computation cost.

Contextual Feature Extractor of k-nearest neighboring regions performs best. To explore the benefits of similarity graph in Contextual Feature Extractor in our framework, we replace the similarity graph in the CCI shown in Fig. 4 (c) with 1) "FC", the fully-connected layer, 2) "MAX", max-pooling layer, 3) "AVE", average-pooling layer after concatenating all the feature vectors of k neighboring regions. The results are shown in Tab. 4. The similarity graph operation can improve all the evaluation metrics compared with the simple fully-connected/ max-pooling/ average-pooling operation after concatenating all the feature vectors of k neighboring regions. That's because the similarity graph not only includes the visual features of k neighboring regions but also utilizes the relation between the target region and the neighboring region. Furthermore, Tab. 4 shows that the nearest-neighbor regions ("Nearest") perform better than the regions randomly-selected from all the regions in the image ("Random") due to more correlated regions involved in the description generation.

k		10	20	30	50	100
mAP	RPN	8.915	9.144	9.109	8.804	8.749
	GT	33.260	33.412	33.411	33.389	33.089

Table 5. Results with different numbers of k-nearest regions for neighboring features in the Contextual Feature Extractor. The results are reported when hyper-parameter k is set as $10, 20, 30, 50, 100$ respectively.

Contextual Feature Extractor with hyper-parameter $k = 20$ outperforms others. The number of neighboring regions is worth investigating because it can be used to find a trade-off between the effective message passing and the noises from non-correlated proposals in the image. We validate the number of neighboring regions among $10, 20, 30, 50$ and 100 of CCI. The results are reported in Tab. 5. We adopt k as 20 for further experiments for the best performance (9.144%) on mAP considering region localization and description jointly.

6. Conclusion

In this paper, we propose a novel end-to-end framework for dense captioning, named as Context and Attribute Grounded Dense Captioning (CAG-Net) by utilizing the visual information of both the target region and multi-scale contextual cues, *i.e.*, *global* and *neighboring*. The proposed contextual feature extractor exploits the message passing between the target region and k-nearest neighboring regions in the image while the attribute grounded contextual cue integration modules reinforce rich and accurate description generation. To enhance the description generation for the regions, we extract linguistic attributes from the reference sentences as the auxiliary supervision at each stage during the training process. Extensive experiments demonstrate the respective effectiveness and significance of the proposed CAG-Net on the challenging large-scale VG dataset.

Acknowledgment This work is supported in part by the National Natural Science Foundation of China (Grant No. 61371192), the Key Laboratory Foundation of the Chinese Academy of Sciences (CXJJ-17S044) and the Fundamental Research Funds for the Central Universities (WK2100330002, WK3480000005), in part by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK14203015, CUHK14239816, CUHK419412, CUHK14207-814, CUHK14208417, CUHK14202217), the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *CVPR*, June 2018.

[3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[4] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, July 2017.

[6] Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[7] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. "factual" or "emotional": Stylized image captioning with adaptive learning and attention. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. In *IJCV*, volume 95, pages 1–12, 2011.

[10] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *ECCV*, 2018.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, volume 88, pages 303–338. Springer, 2010.

[12] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, July 2017.

[13] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with R* CNN. In *ICCV*, 2015.

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[15] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*, 2018.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press, 1997.

[17] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *CVPR*, 2018.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[19] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[20] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

[21] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, volume 123, pages 32–73, 2017.

[22] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 2016.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[24] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[25] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, July 2017.

[27] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, 2018.

[28] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[30] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, July 2017.

[31] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European conference on artificial intelligence*, 2004.

[32] Luming Tang, Yexiang Xue, Di Chen, and Carla P Gomes. Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *AAAI*, 2018.

[33] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV*, 2018.

[34] L. Yang, K. Tang, J. Yang, and L.-J. Li. Dense captioning with joint inference and visual context. In *CVPR*, 2017.

[35] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[36] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017.

[37] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, 2018.

[38] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[39] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, June 2018.