This ICCV 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.

LogicSeg: Parsing Visual Semantics with Neural Logic Learning and Reasoning

Liulei Li[1,2], Wenguan Wang[1]*, Yang Yi[1]
[1] ReLER, CCAI, Zhejiang University [2] ReLER, AAII, University of Technology Sydney
https://github.com/lingorX/LogicSeg/

Abstract

Current high-performance semantic segmentation models are purely data-driven sub-symbolic approaches and blind to the structured nature of the visual world. This is in stark contrast to human cognition which abstracts visual perceptions at multiple levels and conducts symbolic reasoning with such structured abstraction. To fill these fundamental gaps, we devise LogicSeg*, a holistic visual semantic parser that integrates neural inductive learning and logic reasoning with both rich data and symbolic knowledge. In particular, the semantic concepts of interest are structured as a hierarchy, from which a set of constraints are derived for describing the symbolic relations and formalized as first-order logic rules. After fuzzy logic-based continuous relaxation, logical formulae are grounded onto data and neural computational graphs, hence enabling logic-induced network training. During inference, logical constraints are packaged into an iterative process and injected into the network in a form of several matrix multiplications, so as to achieve hierarchy-coherent prediction with logic reasoning. These designs together make* LogicSeg *a general and compact neural-logic machine that is readily integrated into existing segmentation models. Extensive experiments over four datasets with various segmentation models and backbones verify the effectiveness and generality of* LogicSeg. *We believe this study opens a new avenue for visual semantic parsing.*

1. Introduction

Interpreting high-level semantic concepts of visual stimuli is an integral aspect of human perception and cognition, and has been a subject of interest in computer vision for nearly as long as this discipline has existed. As an exemplar task of visual semantic interpretation, *semantic segmentation* aims to group pixels into different semantic units. Progress in this field has been notable since the seminal work of fully convolution networks (FCNs)[1] and been further advanced by the recent launch of fully attention networks (Transformer) [2].

Despite these technological strides, we still observe current prevalent segmentation systems lack in-depth reflection

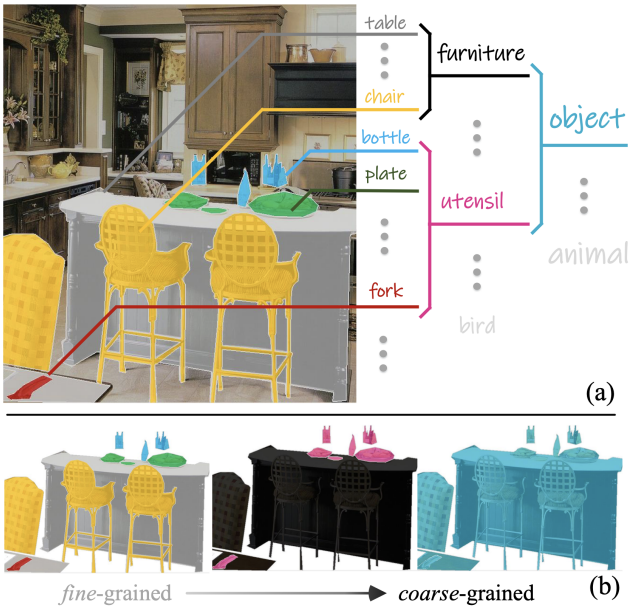


Figure 1: (a) We humans abstract our perception in a structured manner, and conduct reasoning through symbol manipulation over such multi-level abstraction. (b) We aim to *holistically* interpret visual semantics, through the integration of both data-driven sub-symbolic learning and symbolic knowledge-based logic reasoning.

on some intrinsic nature of human cognition. **First**, standard segmentation systems simply assume the semantic concepts in the set of interest have no underlying relation and predict all these concepts *exclusively*. By contrast, humans interpret a scene by components. For example in Fig. 1, we can effortlessly recognize many pieces of `furniture`, such as `chairs` and `tables`, and identify various `utensils`, *e.g.*, `bottles`, and `plates`. Such capacity of structured understanding of visual semantics is an innate aspect of human perception [3], complies with our way of the organization of knowledge [4, 5], and has a close relation to many meta-cognitive skills including *compositional generalization* (*i.e.*, making infinite use of finite means) [6], *systematicity* (*i.e.*, cognitive capacity comes in groups of related behaviours) [7], and *interpretability* (*i.e.*, interpreting complex concepts with simpler ones) [8, 9]. Despite its significance and ubiquity, surprisingly little has been done on the computational mod-

[1]Corresponding author: Wenguan Wang.

eling of structured visual perception in the segmentation literature. Though exceptions exist [10–14], in general they are scattered, lacking systematic study. **Second**, the latest semantic segmentation systems, label structure aware or not, have developed a pure sub-symbolic learning approach. They enjoy the advantages of robust distributed representation of concept entities, but struggle with explicit reasoning with the relations among entities by discrete symbolic representations [15]. Nevertheless, studies in cognition suggest that our perception works at multiple levels of semantic abstraction [16], intertwined with logical reasoning through manipulation of symbolic knowledge/concepts [17]. For example, after recognizing many `utensils` from Fig. 1, we know the scene is more likely a `kitchen`, rather than a `bathroom` or `gym`. This judgement comes as a result of reasoning with some abstract knowledge, such as "*utensils typically appear in the kitchen*" and "*utensils are seldom seen in the bathroom*," which are generalized from our daily experience. The judgement of the scene type may become a belief and in turn cause reallocation of our visual attention [18], hence driving us to find out more relevant details, such as small `forks`.

Filling the gaps identified above calls for a fundamental paradigm shift: **i)** moving away from pixel-wise 'flat' classification towards semantic structure-aware parsing; and **ii)** moving away from the extreme of pure distributed representation learning towards an ambitious hybrid which combines both powerful sub-symbolic learning and principled symbolic reasoning. To embrace this change, we develop LOGICSEG, a structured visual parser which exploits neural computing and symbolic logic in a neural-symbolic framework for holistic visual semantic learning and reasoning. In particular, given a set of hierarchically-organized semantic concepts as background knowledge and parsing target, we first use *first-order logic*, a powerful declarative language, to comprehensively specify relations among semantic classes. After *fuzzy logic* based relaxation, the logical formulae of hierarchy constraints can be grounded on data. During training, each logical constraint is converted into a differentiable loss function for gradient descent optimization. During inference, the logical constraints are involved into an iterative process, and calculated in matrix form. This not only ensures the observance of the compositional semantic structure but also binds logic reasoning into network feed-forward prediction.

By accommodating logic-based symbolic rules into network training and inference, our LOGICSEG **i)** blends statistical learning with symbolic reasoning, **ii)** obtains better performance, and **iii)** guarantees its parsing behavior compliant with the logically specified symbolic knowledge. We also remark that our study is relevant to a field of research called *neural-symbolic computing* (NSC) [19–21]. With the promise of integrating two critical cognitive abilities [22]: inductive learning (*i.e.*, the ability to learn general principles from experience) and deductive reasoning (*i.e.*, the ability to draw logical conclusions from what has been learned), NSC has long been a multi-disciplinary research focus and shown superiority in certain application scenarios, such as program generation [23–25], and question answering [26, 27]. This work unlocks the potential of NSC in visual semantic parsing – a fundamental, challenging, and large-scale vision task.

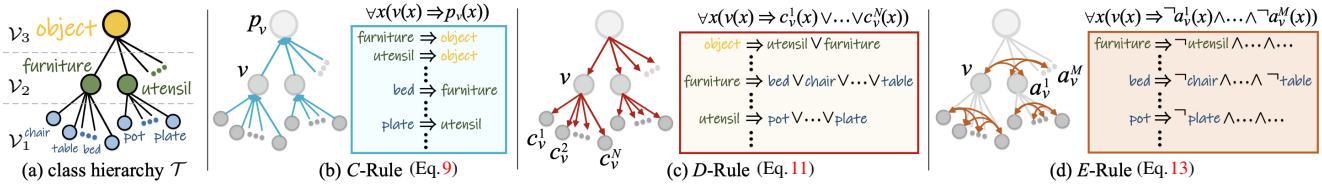
LOGICSEG is a principled framework. It is fully compatible with existing segmentation network architectures, with only minor modification to the classification head and a plug-and-play logic-induced inference module. We perform experiments on four datasets covering wide application scenarios, including automated-driving (Mapillary Vistas 2.0 [28], Cityscapes [29]), object-centric (Pascal-Part [30]), and daily (ADE-20K [31]) scenes. Experimental results show that, on the top of various segmentation models (*i.e.*, DeepLabV3+ [32], Mask-2Former [33]) and backbones (*i.e.*, ResNet-101 [34], Swin-T [35]), LOGICSEG yields solid performance gains (**1.12%**-**3.29%** mIoU) and suppresses prior structured alternatives. The strong generalization and promising performance of LOGICSEG evidence the great potential of integrating symbolic reasoning and sub-symbolic learning in machine perception.

2. Related Work

Semantic Segmentation. Since the proposal of fully convolutional networks (FCNs) [1], research studies in pixel-level semantic interpretation have witnessed a phenomenal growth. Tremendous progress has been achieved by, for example, polishing context cues [36–52], investigating boundary information [53–58], incorporating neural attention [59–71], adopting data structure-aware learning [72–76], and automating network engineering [77–80]. More recently, the engagement of advanced Transformer [2] architecture, which specializes in long-range dependency modeling, is widely viewed as a promising route for further development [33, 81–86].

Though impressive, existing segmentation solutions are mainly aware of straightforward prediction for *flatten* labels. They are largely blind to the rich structures among semantic concepts and lack an explicit mechanism for symbol manipulation/logical calculus, which is what distinguishes humans from other animals [87**?** , 88]. This work represents a small yet solid step towards addressing these fundamental limitations through an integrated neural-logic machine, and inspects semantic segmentation from a brand-new standpoint.

Label Structure-aware Semantic Segmentation. Till now, only a rather small number of deep learning based segmentation models [10, 13, 89–91] are built with structured label taxonomies. The origin of this line of research can be traced back to the task of *image parsing* [89, 90, 92–96] raised in the pre-deep learning era. Basically, image parsing seeks for a holistic explanation of visual observation: scenes can be understood as a sum of novel objects, and the objects can be further broken down into fine-grained parts. In the deep



(a) class hierarchy $\mathcal{T}$ (b) C-Rule (Eq. 9) (c) D-Rule (Eq. 11) (d) E-Rule (Eq. 13)

Figure 2: Illustration of the (a) class hierarchy $\mathcal{T}$, and (b-d) abstract relational knowledge specified by first-order logic formulae (§3.1).

learning era, the majority of structured segmentation models are dedicated to *human parsing* [89, 90, 97, 98], which is customized to human-part relation understanding. As for the case of general-purpose segmentation, there are far rare literature [10–13, 91], and many of them incorporate label taxonomies into the network topology, losing generality [10–12]. As a notable exception, [13] converts the task as *pixel-wise multi-label classification* and exploits the class hierarchy for training regularization, with only trivial architectural change.

In a nutshell, previous efforts highlight the limits of standard segmentation models for semantic structures. However, they typically **i)** resolve to stand on the side of sub-symbolic learning, **ii)** make usage of only fragments of structured relations (for instance, the exclusion relation is neglected by [13]), **iii)** lack structure-ware inference, and/or **iv)** rely on sophisticated and specialized neural structures. By contrast, we formulate the structured task into a neural-symbolic framework. We derive a comprehensive set of symbolic relational knowledge in the form of first-order logic and deeply embed logical constraints into network training and inference. Our algorithm is a general framework that is applicable to existing standard hierarchy-agnostic segmentation architectures.

Neuro-Symbolic Computing. There has been a line of research, called neural-symbolic computing (NSC), that pursues the integration of the symbolic and statistical paradigms of cognition [19–21]. NSC has a long history, dating back to McCulloch and Pitts 1943 paper [99], even before AI was recognized as a new scientific field. During 2000s, NSC received systematic study [100–103]. Early NSC systems were meticulously designed for hard logic reasoning, but they are far less trainable, and fall short when solving real-world problems. NSC has recently ushered in its renaissance, since it shows promise of reconciling statistical learning of neural networks and logic reasoning of abstract knowledge – which is viewed as a key enable to the next generation of AI [104, 105]. Specifically, recent NSC systems [106, 107] show the possibility for modern neural networks to manipulate abstract knowledge with diverse forms of symbolic representation, including knowledge graph [108–110], propositional logic [111–113], and first-order logic [114–116]. They also demonstrate successful application in several domains and disciplines, *e.g.*, scientific discovery [117, 118], program generation [23–25], (visual) question-answering [26, 27], robot planning [119–121], and mathematical reasoning [122–124].

To date, none of NSC systems reports advanced performance in large-scale vision, to our best knowledge. In this work, we take the lead to promote and implement the idea of conciliating the methodologies of symbolic and neural paradigms, in visual semantic interpretation. Moreover, many previous NSC systems only exploit logical constraints during network training [113, 116, 125–128], while our solution is more favored as logic rules are involved throughout network training and inference. As a result, impressive performances across diverse challenging datasets are delivered, and in turn, provide solid empirical evidence for the power of NSC.

3. Methodology

Task Setup and Notations. In this work we are interested in structured visual parsing [13] – a more challenging yet realistic setting for semantic segmentation – where both semantic concepts and their relations are considered in a form of a tree-shaped class hierarchy $\mathcal{T} = \langle \mathcal{V}, \mathcal{E} \rangle$. The node set $\mathcal{V} = \cup_{l=1}^{L} \mathcal{V}_l$ represents the classes/concepts at L abstraction levels. For instance in Fig. 2(a), the leaf nodes $\mathcal{V}_1$ are the finest classes (*e.g.*, `chair`, `pot`), while the internal nodes are higher-level concepts (*e.g.*, `furniture`, `utensil`), and the roots $\mathcal{V}_L$ are the most abstract ones (*e.g.*, `object`). The edge set $\mathcal{E}$ encodes relational knowledge among classes. For example, a directed edge $u \to v \in \mathcal{E}$ denotes a *part-of* relation between classes $u, v \in \mathcal{V}$ in *adjacent* levels (*e.g.*, `utensil`→`pot`).

Given $\mathcal{T}$, the target goal is to assign each pixel a *valid* root-to-leaf path in $\mathcal{T}$. For instance, associating a pixel with `object`→`utensil`→`pot` is valid, yet with `object`→`furniture`→`pot` is *invalid*. Thus standard semantic segmentation can be viewed as a specific case of such structured setting — only assigning pixels with one single class label from the leaf nodes $\mathcal{V}_1$ without considering the hierarchy.

Algorithmic Overview. LogicSeg is a unified, neural-logic learning and reasoning model for visual parsing, supported by large-scale data and the structured symbolic knowledge $\mathcal{T}$.

- From the ***neural*** aspect, LogicSeg is *model-agnostic*. After dense feature extraction, its classification head outputs a total of $|\mathcal{V}|$ *sigmoid*-normalized scores, *i.e.*, $\boldsymbol{s} \in [0,1]^{|\mathcal{V}|}$, over all the classes $\mathcal{V}$ for each pixel, like [13]. Here $|\cdot|$ counts its elements. A set of logic rules, derived from $\mathcal{T}$, are injected into network training and inference.
- From the ***logic*** aspect, LogicSeg uses *first-order logic* to express the complex and abstract relational knowledge in $\mathcal{T}$. The network is learnt as approximation of logic predicates by following the logical specifications. Once trained, it conducts iterative reasoning on the basis of logic rules.

After introducing our logic based visual relational knowledge

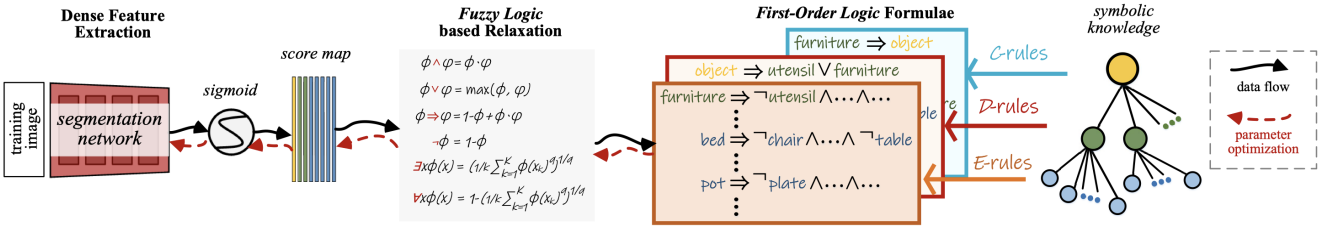


Figure 3: Illustration of our logic-induced network training (§3.2). For clarity, the pixel-wise binary cross-entropy loss is omitted.

representation (§3.1), we will elaborate on our logic-induced network training (§3.2) and inference (§3.3) strategies.

3.1. Parsing Visual Semantics with Logic Rules

We formalize our target task — *learning and reasoning visual semantics with logic* — as a triple $\langle \mathcal{T}, \mathcal{X}, \Pi \rangle$. $\mathcal{X}$ is a data collection, *i.e.*, $\mathcal{X} = \{(x_k, \boldsymbol{y}_k)\}_{k=1}^{K}$, where x_k is a pixel data point, and $\boldsymbol{y}_k \in \{0,1\}^{|\mathcal{V}|}$ is its groundtruth symbolic description in terms of the semantic hierarchy $\mathcal{T}$. Π is a set of hierarchy rules declaratively expressed by *first-order logic*, containing **i)** *constants*, *e.g.*, pixel samples $x_1, x_2, \cdots$; **ii)** *variables* ranging over constants, *e.g.*, x; and **iii)** *unary predicates*, one for each class $v \in \mathcal{V}$, denote the semantics of variables and return *true* and *false*, *e.g.*, $\texttt{bed}(x) = \textit{true}$ states the fact that pixel x belongs to a bed. A logic rule/formula is a sequence of finite predicates with *connectives* (*i.e.*, $\wedge, \vee, \neg, \Rightarrow$) and *quantifiers* (*i.e.*, $\forall, \exists$), and organized in *prenex* form in our case.

Concretely, Π is composed of three types of rules, *i.e.*, *composition*, *decomposition*, and *exclusion*, for comprehensively describing the structured symbolic knowledge $\mathcal{T}$.

• ***Composition Rule*** (*C*-rule) expresses our knowledge about the *composition* relations between semantic concepts, such as "*bed and chair are (subclasses of) furniture*," in a form of:

$$\begin{aligned} \forall x(\texttt{bed}(x) &\Rightarrow \texttt{furniture}(x)), \\ \forall x(\texttt{chair}(x) &\Rightarrow \texttt{furniture}(x)), \end{aligned} \tag{1}$$

where `bed`, `chair`, `furniture` are predicates, and '$\phi \Rightarrow \varphi$' indicates φ is a logical consequence of antecedence ϕ.

Definition 3.1.1 (*C*-rule)**.** *If one class is labeled true, its superclass should be labeled true* (Fig. 2(b)):

$$\forall x(v(x) \Rightarrow p_v(x)), \tag{2}$$

where p_v is the parent node of v in $\mathcal{T}$, *i.e.*, $p_v \rightarrow v \in \mathcal{E}$ (the tree structure of $\mathcal{T}$ restricts each class to possess only one superclass). *C*-rule generalizes the famous *tree-property* [129, 130].

• ***Decomposition Rule*** (*D*-rule) states our knowledge about the *decomposition* relations among semantic concepts, such as "*furniture is the superclass of bed, chair, $\cdots$, table*," via:

$$\begin{aligned} \forall x(\texttt{furniture}(x) \Rightarrow &\texttt{bed}(x) \vee \texttt{chair}(x) \vee \\ &\cdots \vee \texttt{tabel}(x)). \end{aligned} \tag{3}$$

Definition 3.1.2 (*D*-rule)**.** *If one class is labeled true, at least one of its subclasses should be labeled true* (Fig. 2(c)):

$$\forall x(v(x) \Rightarrow c_v^1(x) \vee c_v^2(x) \vee \cdots \vee c_v^N(x)), \tag{4}$$

where $c_v^n \in \mathcal{C}_v$ are all the child nodes of v in $\mathcal{T}$, *i.e.*, $v \rightarrow c_v^n \in \mathcal{E}$. *C*-rule and *D*-rule are not equivalent. For instance in Eq. 1, $\texttt{bed}(x)$ is sufficient but not necessary for $\texttt{furniture}(x)$: given the fact "x is furniture", we cannot conclude "x is bed".

• ***Exclusion Rule*** (*E*-rule) specifies our knowledge about *mutual exclusion* relations between *sibling* concepts, such as "*a bed cannot be at the same time a chair*," in a form of:

$$\forall x(\texttt{bed}(x) \Rightarrow \neg \texttt{chair}(x)). \tag{5}$$

Definition 3.1.3 (*E*-rule)**.** *If one class is labeled true, all its sibling classes should be labeled false* (Fig. 2(d)):

$$\forall x(v(x) \Rightarrow \neg a_v^1(x) \wedge \neg a_v^2(x) \wedge \cdots \wedge \neg a_v^M(x)), \tag{6}$$

where $a_v^m \in \mathcal{A}_v$ are all the peer nodes of v in $\mathcal{T}$. Note that *E*-rule is ignored by many hierarchy-aware algorithms [13, 131, 132].

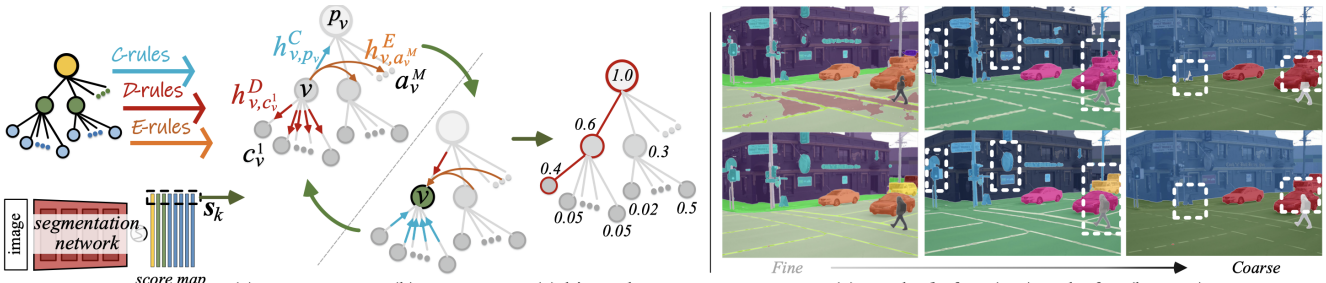
3.2. Logic-Induced Training

So far, we shown the logic rules Π provide LogicSeg a flexible language for comprehensively expressing the complex *meronymy* and *exclusion* relations among symbolic concepts in the hierarchy $\mathcal{T}$. Unfortunately, these rules are logic formulae working with variables (assuming a boolean value), and non-differentiable logic symbols (*e.g.*, $\forall, \Rightarrow$). This prevents the integration with end-to-end network learning.

Inspired by [128, 133], a *fuzzy logic* based *grounding* process is adopted to interpret logic formulae as differentiable fuzzy relations on real numbers for neural computing (Fig. 3).

Fuzzy relaxation. Fuzzy logic is a form of soft probabilistic logic. It deals with reasoning that is approximate instead of fixed and exact; variables have a truth degree that ranges in $[0, 1]$: zero and one meaning that the variable is *false* and *true* with certainty, respectively [134]. Hence we can ground predicates onto segmentation network outputs. For instance, given a pixel sample x, corresponding network prediction score w.r.t. class *bed* is a grounded predicate w.r.t. $\texttt{bed}(x)$. Logical connectives, *i.e.*, $\wedge, \vee, \neg, \Rightarrow$ are approximated with *fuzzy operators*, *i.e.*, *t-norm*, *t-conorm*, *fuzzy negation*, and *fuzzy implication*. As suggested by [133], we adopt the operators in *Goguen fuzzy logic* [135] and *Gödel fuzzy logic* [136]:

$$\begin{aligned} \phi \wedge \varphi &= \phi \cdot \varphi, \quad & \phi \vee \varphi &= \max(\phi, \varphi), \\ \neg\phi &= 1 - \phi, \quad & \phi \Rightarrow \varphi &= 1 - \phi + \phi \cdot \varphi. \end{aligned} \tag{7}$$



(a) message creation (Eq. 16) (b) message aggregation (Eq. 17) (c) hierarchy-aware parsing (Eq. 18) (e) results *before* (top) and *after* (bottom) logic-induced inference

Figure 4: Illustration of our logic-induced inference (§3.3). (a-b) Iterative reasoning is made by exchanging and absorbing messages between nodes, following the logic rules Π. For clarity, we only show the message creation (Eq. 16) and aggregation (Eq. 17) stages for one single node. (c) Structured parsing (Eq. 18) is conducted by selecting the top-scoring path $\mathcal{P}^*$ (highlighted in red) after logic-guided iterative reasoning. (d) With logic-induced inference, LOGICSEG is able to generate more accurate and hierarchy-compliant predictions.

The existential quantifier $\exists$ and universal quantifier $\forall$ are approximated in a form of generalized mean:

$$\begin{aligned}\exists x\phi(x) &= \left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\phi(x_k)^q\right)^{\frac{1}{q}},\\ \forall x\phi(x) &= 1-\left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}(1-\phi(x_k))^q\right)^{\frac{1}{q}},\end{aligned} \tag{8}$$

where $q\in\mathbb{Z}$. Please refer to [128, 133] for detailed discussion regarding the rationale behind such approximation of $\exists$ and $\forall$.

Logic Loss. With fuzzy relaxation, we are ready to convert our first-order logic rules Π into loss functions.

• ***C*-rule Loss.** For a non-root node $v\in\mathcal{V}/\mathcal{V}_L$, its corresponding *C*-rule (*cf.* Eq. 2) is grounded as:

$$\mathcal{G}_C(v)=1-\left(\tfrac{1}{K}\textstyle\sum_{k=1}^{K}(\boldsymbol{s}_k[v]-\boldsymbol{s}_k[v]\cdot\boldsymbol{s}_k[p_v])^q\right)^{\frac{1}{q}}, \tag{9}$$

where $\boldsymbol{s}_k[v]\in[0,1]$ refers to the score (confidence) of x_k for class v. Then the *C*-rule based training objective is given as:

$$\mathcal{L}_C=\tfrac{1}{|\mathcal{V}|-|\mathcal{V}_L|}\textstyle\sum_{v\in\mathcal{V}/\mathcal{V}_L}1-\mathcal{G}_C(v). \tag{10}$$

• ***D*-rule Loss.** For a non-leaf node $v\in\mathcal{V}/\mathcal{V}_1$, its corresponding *D*-rule (*cf.* Eq. 4) is grounded as:

$$\mathcal{G}_D(v)=1-\left(\tfrac{1}{K}\sum_{k=1}^{K}(\boldsymbol{s}_k[v]-\boldsymbol{s}_k[v]\cdot\texttt{max}(\{\boldsymbol{s}_k[c_v^n]\}_n))^q\right)^{\frac{1}{q}}. \tag{11}$$

Similarly, our *D*-rule loss is given as:

$$\mathcal{L}_D=\tfrac{1}{|\mathcal{V}|-|\mathcal{V}_1|}\textstyle\sum_{v\in\mathcal{V}/\mathcal{V}_1}1-\mathcal{G}_D(v). \tag{12}$$

• ***E*-rule Loss.** During the grounding of *E*-rule (*cf.* Eq. 6), we first translate such *one-vs-all* exclusion statement to a semantically equivalent expression, *i.e.*, the aggregation of multiple *one-vs-one* exclusion ($\{(v(x)\Rightarrow\neg a_v^1(x)),\cdots,\{(v(x)\Rightarrow\neg a_v^M(x))\}$). Adopting such translation is to avoid the *sorites paradox*, *i.e.*, a long chain of only slightly unreliable deductions can be very unreliable [137] (*e.g.*, $0.9^{10}\approx0.34$), happened during approximating a series of $\wedge$. Then, for each node $v\in\mathcal{V}$, its corresponding *E*-rule is grounded as:

$$\mathcal{G}_E(v)=1-\tfrac{1}{M}\textstyle\sum_{m=1}^{M}\left(\tfrac{1}{K}\sum_{k=1}^{K}(\boldsymbol{s}_k[v]\cdot\boldsymbol{s}_k[a_v^m])^q\right)^{\frac{1}{q}}. \tag{13}$$

Similarly, our *E*-rule loss is given as:

$$\mathcal{L}_E=\tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}1-\mathcal{G}_E(v). \tag{14}$$

In this way, it is possible to backpropagate the gradient from the logic loss into the network. The network is essentially learned as neural predicates obeying the logical constraints. It is worth mentioning that, due to large-scale training, it is infeasible to compute the full semantics of $\forall$; batch-training can be viewed as sampling based approximation [133].

Our overall training target is organized as:

$$\mathcal{L}=\alpha(\mathcal{L}_C+\mathcal{L}_D+\mathcal{L}_E)+\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\mathcal{L}_{\text{BCE}}(\boldsymbol{s}_k,\boldsymbol{y}_k). \tag{15}$$

Here $\boldsymbol{y}\in\{0,1\}^{|\mathcal{V}|}$ is the groundtruth, $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss, and the coefficient is empirically set as $\alpha=0.2$.

3.3. Logic-Induced Inference

We just showed that LOGICSEG can approximate the predicates by integrating symbolic logic constraints into large-scale network training. However, during inference, there is no explicit way to ensure the alignment between the class hierarchy $\mathcal{T}$ and network prediction, neither sound reasoning with the logic rules Π. We thus put forward *logic-induced reasoning* (Fig. 4), where the logic rules Π are encapsulated into an iterative optimization process. Such process is non-learnable, based on only matrix operations and thus can be seamlessly embedded into network feed-forward inference, yielding an elegant yet compact neural-logic visual parser.

Our solution is built upon the classic *message passing* algorithm which is to estimate the marginal likelihood for a given tree structure by *iteratively* exchanging messages between nodes. Specifically, at each iteration, for each pixel sample x_k, node $v\in\mathcal{V}$ sends different types of messages to different neighboring nodes, according to the logic rules Π:

$$\begin{aligned}&C\text{-message: } h^C_{v,p_v}=v(x_k)\Rightarrow p_v(x_k)\\ &\qquad=1-\boldsymbol{s}_k[v]+\boldsymbol{s}_k[v]\cdot\boldsymbol{s}_k[p_v],\\ &D\text{-message: } h^D_{v,c_v^n}=v(x_k)\Rightarrow c_v^1(x_k)\vee\cdots\vee c_v^N(x_k)\\ &\qquad=1-\boldsymbol{s}_k[v]+\boldsymbol{s}_k[v]\cdot\texttt{max}(\{\boldsymbol{s}_k[c_v^n]\}_n),\\ &E\text{-message: } h^E_{v,a_v^m}=-1\cdot\left(v(x_k)\Rightarrow\neg a_v^1(x_k)\wedge\cdots\wedge\neg a_v^M(x_k)\right)\\ &\qquad=-\left(1-\tfrac{1}{M}\textstyle\sum_{m=1}^{M}\boldsymbol{s}_k[v]\cdot\boldsymbol{s}_k[a_v^m]\right).\end{aligned} \tag{16}$$

Node v is updated by aggregating the received messages:

$$\boldsymbol{s}_k[v] \leftarrow \boldsymbol{s}_k[v] + \tfrac{1}{N}\textstyle\sum_{c_v^n \in \mathcal{C}_v} \boldsymbol{s}_k[c_v^n] \cdot h^C_{c_v^n, v} + \boldsymbol{s}_k[p_v] \cdot h^D_{p_v, v} + \tfrac{1}{M}\textstyle\sum_{a_v^m \in \mathcal{A}_v} \boldsymbol{s}_k[a_v^m] \cdot h^E_{a_v^m, v}. \quad (17)$$

Each message (*cf.* Eq. 16) accounts for the certainty degree that v satisfies the corresponding logic rule (*cf.* §3.1) when being grounded on pixel data point x_k, with fuzzy logic based approximation (*cf.* §3.2). Intuitively, the more certainty a node meets the logic rules, the more message it can propagate to other nodes. Note that, v creates a *negative* message h^E_{v,a_v^m} to "suppress" other peer nodes due to their exclusion relations. In Eq. 17, the received messages are weighted by the confidence of the source nodes themselves – the grounded predicates, *i.e.*, $\boldsymbol{s}_k[c_v^n]$, $\boldsymbol{s}_k[p_v]$, and $\boldsymbol{s}_k[a_v^m]$. After each iteration, the score vector $\boldsymbol{s}_k$ is *softmax*-normalized per hierarchy level.

Finally, each pixel x_k is associated with the top-scoring *root-to-leaf* path in the hierarchy $\mathcal{T}$ (red path in Fig. 4(c)):

$$\mathcal{P}^* = \{v_1^*, \cdots, v_L^*\} = \underset{\mathcal{P} \subset \mathcal{T}}{\operatorname{argmax}} \textstyle\sum_{v^{\mathcal{P}} \in \mathcal{P}} \boldsymbol{s}_k[v^{\mathcal{P}}], \quad (18)$$

where $\mathcal{P} = \{v_1^{\mathcal{P}}, \cdots, v_L^{\mathcal{P}}\} \subset \mathcal{T}$ indicates a feasible root-to-leaf path in $\mathcal{T}$, *i.e.*, $\forall v_l^{\mathcal{P}}, v_{l-1}^{\mathcal{P}} \in \mathcal{P} \Rightarrow v_l^{\mathcal{P}} \rightarrow v_{l-1}^{\mathcal{P}} \in \mathcal{E}$.

It is easy to find that all the logic-induced inference steps (*cf.* Eq. 16-18) can be formulated in *matrix* form with only a couple of matrix multiplications (see corresponding pseudo-code in the supplementary). Hence it is efficient on GPU and can be straightforward injected into the network, making LogicSeg a fully-integrated neural-logic machine. In practice, 2-iteration message passing is enough for robust prediction. Through logic-induced reasoning (*cf.* Eq. 17) and hierarchy-aware parsing (*cf.* Eq. 18), LogicSeg is able to **i)** obtain *improved performance*, and **ii)** guarantee the parsing results to *respect the hierarchy* $\mathcal{T}$, with **iii)** only *negligible speed delay* (about 3.8%). See §4.4 for related experiments.

4. Experiment

4.1. Experimental Setup

Datasets. We conduct extensive experiments on four datasets, *i.e.*, Mapillary Vistas 2.0 [28], Cityscapes [29], Pascal-Part-108 [30], and ADE20K [31]. The four datasets are selected to cover the rich application scenarios of semantic segmentation, including urban street segmentation for automated driving (*i.e.*, [28, 29]), object part parsing (*i.e.*, [30]), and fine-grained understanding of daily scenes (*i.e.*, [31]), so as to comprehensively examine the utility of our algorithm.

- **Mapillary Vistas 2.0** is a large-scale urban scene dataset. It contains $18,000/2,000/5,000$ images for `train/val/test`. A three-level semantic hierarchy, covering $4/16/124$ concepts, is officially provided for dense annotation.
- **Cityscapes** has $2,975/500/1,524$ finely annotated, urban street images for `train/val/test`. The label hierarchy consists of 19 fine-gained concepts and 6 superclasses.
- **Pascal-Part-108** is the largest object part parsing dataset. It consists of $4,998/5,105$ images for `train/test`. To establish the class hierarchy, we group 108 part-level labels into 20 object-level categories, as in [138–141].
- **ADE20K** is a large-scale generic scene parsing dataset. It is divided into $20,210/2,000/3,000$ images for `train/val/test`. It provides pixel-wise annotations for 150 fine-grained semantic classes, from which a three-layer label hierarchy (with 3/14/150 concepts) can be derived.

Evaluation Metric. We adopt the standard metric, mean intersection-over-union (mIoU), for evaluation. For detailed performance analysis, the score is reported for each hierarchy level l (denoted as mIoU^l), as suggested by [13, 89].

Base Models and Competitors. To demonstrate our wide benefit, we approach our algorithm on two famous segmentation architectures, *i.e.*, DeepLabV3+ [32] and Mask2Former [33], with ResNet-101 [34] and Swin-T [35] backbones. For performance comparison, we involve several hierarchy-aware segmentation models [13, 138, 141], and view HSSN [13] as our major rival as it is a general framework that reports strong results over several datasets, instead of the others that are dedicated to specific dataset(s) or task setup(s). For comprehensive evaluations, we include a group of previous hierarchy-agnostic segmentation algorithms [10, 38, 44, 81–83], whose segmentation results on coarse-grained semantics are obtained by merging the predictions of the corresponding subclasses.

Training. For the sake of fairness, we follow the standard training setup in [44, 74, 84, 142, 143]. In particular, we train 240K/80K iterations for Mapillary Vistas 2.0/Cityscapes, with batch size 8 and crop size 512×1024; 60K/160K iterations for Pascal-Part-108/ADE20K, with batch size 16 and crop size 512×512. For data augmentation, the images are horizontally flipped and scaled with a ratio between 0.5 and 2.0 at random. For network optimization, SGD (with initial learning rate 1e-2, momentum 0.9, and weight decay 1e-4) and Adam (with initial learning rate 6e-5 and weight decay 0.01) are respectively used for CNN-based and neural attention-based models, where the learning rate is scheduled by the polynomial annealing rule. For network initialization, ImageNet [144] pre-trained weights are pre-loaded.

Testing. For Mapillary Vistas 2.0 and Cityscapes, we keep the original image aspect ratio but resize the short edge to 1024. Sliding window inference with the identical window shape as the training size is adopted to save memory. For ADE20K and Pascal-Part-108, the short edge is resized to 512 so as to enable one-time inference for the whole image. As in [44, 67, 84, 98], performance of all the models is reported at multiple scales ($\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$) with horizontal flipping.

Hyperparameters. We set $\alpha = 0.2$ for the loss coefficient (*cf.* Eq. 15), and $q = 5$ for logic quantifier approximation (*cf.* Eq. 8), as suggested by [128]. For network inference, we find 2 iterations of message passing are enough.

Method	Backbone	mIoU^3↑	mIoU^2↑	mIoU^1↑
Seamless [145]	ResNet-101	84.23	70.24	38.82
OCRNet [44]	HRNet-W48	84.19	69.82	38.26
HMSANet [90]	HRNet-W48	84.63	70.71	39.53
Mask2Former [33]	Swin-S	88.81	74.98	43.49
HSSN [13]	ResNet-101	85.27	71.40	40.16
HSSN [13]	Swin-S	90.02	75.81	43.97
DeepLabV3+ [32]	ResNet-101	81.86	68.17	37.43
+ LOGICSEG		**85.51** ↑3.65	**71.69** ↑3.42	**40.72** ↑3.29
MaskFormer [84]	Swin-S	87.93	73.88	42.16
+ LOGICSEG		**90.35** ↑2.42	**76.61** ↑2.73	**45.12** ↑2.96

Table 1: **Urban scene parsing results** (§4.2) on MapillaryVistas 2.0 [28] `val` with a three-level label hierarchy of 4/16/124 concepts.

Method	Backbone	mIoU^2↑	mIoU^1↑
PSPNet [146]	ResNet-101	91.67	80.91
DANet [66]	ResNet-101	91.83	81.52
CCNet [67]	ResNet-101	91.70	81.08
OCRNet [44]	HRNet-W48	92.57	82.33
SETR [83]	ViT-L	92.86	82.75
SegMentor [82]	ViT-L	91.79	81.30
UperNet [10]	Swin-S	91.92	81.79
Mask2Former [33]	Swin-S	93.68	83.62
SegFormer [81]	MiT-B4	93.81	83.90
HSSN [13]	ResNet-101	93.31	83.02
HSSN [13]	Swin-S	94.39	83.74
DeepLabV3+ [32]	ResNet-101	92.16	82.08
+ LOGICSEG		**93.37** ↑1.21	**83.20** ↑1.12
MaskFormer [84]	Swin-S	92.96	82.57
+ LOGICSEG		**94.31** ↑1.35	**83.85** ↑1.28

Table 2: **Urban scene parsing results** (§4.2) on Cityscapes [29] `val` with a two-level label hierarchy of 6/19 concepts.

4.2. Quantitative Comparison Result

MapillaryVistas 2.0 [28] `val`. From Table 1 we can observe that our approach provides notable performance gains over the baselines. For example, our algorithm promotes classic DeepLabV3+ [32] by **3.65%/3.42%/3.29%** over the three semantic levels. On top of MaskFormer [84], our algorithm further lifts the scores by **2.42%/2.73%/2.96%**, suppressing previous hierarchy-agnostic models, as well as HSSN [13] – a newly proposed hierarchy-aware segmentation model.

Cityscapes [29] `val`. Table 2 confirms again our compelling performance in challenging urban street scenes and wide benefits for different segmentation models, *i.e.*, **1.21%/1.12%** over DeepLabV3+, and **1.35%/1.28%** over MaskFormer. Though both encoding concept structures into segmentation, our algorithm greatly outperforms HSSN, suggesting the superiority of our logic reasoning framework.

Pascal-Part-108 [30] `test`. As illustrated by Table 3, our algorithms yields remarkable performance on explaining the compositionality of object-centric semantic structures. Specifically, our algorithm not only consistently boosts the performance of base segmentation models [32, 33], but also defeats two outstanding hierarchy-agnostic competitors [38, 141] as well as three structured alternatives [13, 138, 141].

Method	Backbone	mIoU^2↑	mIoU^1↑
SegNet [38]	ResNet-101	59.81	36.42
FCN-8s [1]	ResNet-101	62.26	38.62
BSANet [140]	ResNet-101	69.37	47.36
GMNet [138]	ResNet-101	69.28	47.21
FLOAT [141]	ResNet-101	70.03	48.08
HSSN [13]	ResNet-101	72.91	48.32
HSSN [13]	Swin-S	77.01	54.79
DeepLabV3+ [32]	ResNet-101	70.86	46.54
+ LOGICSEG		**73.68** ↑2.82	**49.13** ↑2.69
MaskFormer [84]	Swin-S	75.78	53.07
+ LOGICSEG		**77.92** ↑2.14	**55.53** ↑2.46

Table 3: **Object part parsing results** (§4.2) on PASCAL-Part-108 [30] `test` with a two-level label hierarchy of 20/108 concepts.

Method	Backbone	mIoU^3↑	mIoU^2↑	mIoU^1↑
OCRNet [44]	HRNet-W48	76.33	55.76	44.92
SETR [83]	ViT-L	78.92	59.03	49.41
UperNet [10]	Swin-S	78.90	59.17	49.47
SegMentor [82]	ViT-S	77.32	57.18	46.82
K-Net [147]	Swin-S	79.11	59.38	49.95
SegFormer [81]	MiT-B4	79.85	60.24	51.08
GMMSeg [74]	MiT-B5	80.13	60.91	52.12
Mask2Former [33]	Swin-S	80.46	61.15	52.43
HSSN [13]	ResNet-101	79.23	58.52	47.69
HSSN [13]	Swin-S	82.59	62.56	52.37
DeepLabV3+ [32]	ResNet-101	77.24	56.87	46.43
+ LOGICSEG		**79.60** ↑2.36	**59.04** ↑2.17	**48.46** ↑2.03
MaskFormer [84]	Swin-S	79.89	60.32	51.04
+ LOGICSEG		**82.45** ↑2.56	**62.44** ↑2.12	**52.82** ↑1.78

Table 4: **Generic scene parsing results** (§4.2) on ADE20K [31] `val` with a three-level label hierarchy of 3/14/150 concepts.

ADE20K [31] `val`. Table 4 presents our parsing results in general scenes. With a relatively conservative baseline, *i.e.*, DeepLabV3+ [32], our algorithm earn **79.60%**, **59.04%**, and **48.46%**, in terms of mIoU^1, mIoU^2, and mIoU^3 respectively. It delivers a solid overtaking against Mask2Former [33], which is built upon a more advanced architecture. When applied to MaskFormer [84], our algorithm achieves **82.45%/62.44%/52.82%**, pushing forward the state-of-the-art.

Taking together, our extensive benchmarking results provide solid evidence that our algorithm successfully unlocks the power of logic reasoning in large-scale visual parsing, and owns broad applicability across various task scenarios, segmentation architectures, and backbone networks.

4.3. Qualitative Comparison Result

Fig. 5 visualizes qualitative comparisons of LOGICSEG against DeepLabV3+ [32] (*left*) and Mask2Former [33] (*right*) on Mapillary Vistas 2.0 dataset [28]. As seen, with the help of symbolic reasoning, LOGICSEG can generate higher-quality predictions even in challenging scenarios.

4.4. Diagnostic Experiment

For thorough evaluation, we perform a series of ablative studies on Mapillary Vistas 2.0 [28] `val`. All variants are

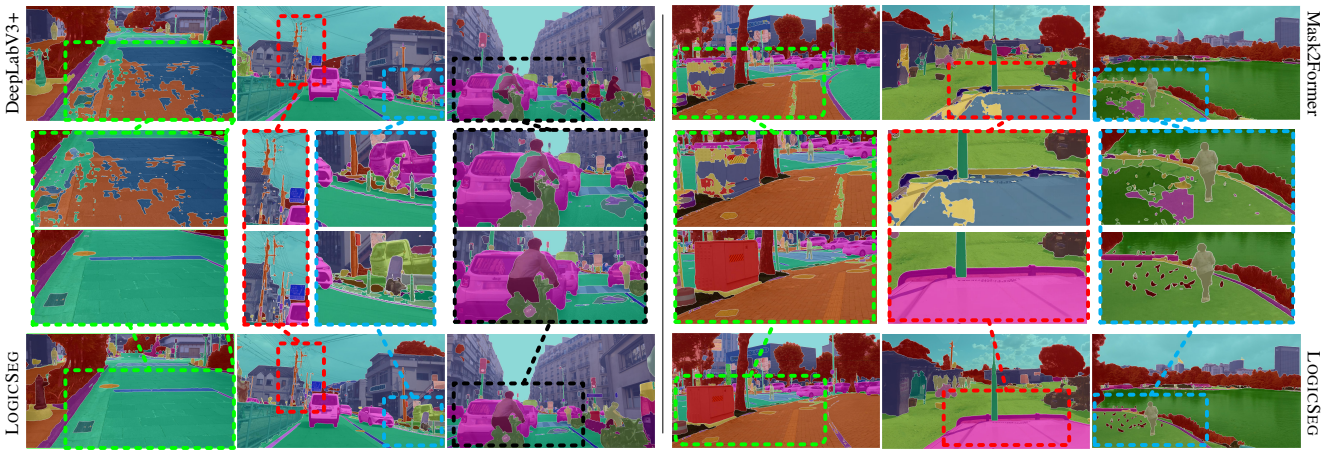


Figure 5: **Visual results** (§4.3) on Mapillary Vistas 2.0 [28]. *Left*: DeepLabV3+ [32] *vs* LogicSeg; *Right*: Mask2Former [33] *vs* LogicSeg.

$\mathcal{L}_C$ Eq.9	$\mathcal{L}_D$ Eq.11	$\mathcal{L}_E$ Eq.13	mIoU^3↑ (multi scale)	mIoU^2↑ (multi scale)	mIoU^1↑ (multi scale)	Training Speed (min/epoch)
			81.86	68.17	37.43	45.62
✓			83.56	69.74	38.71	46.35 +1.60%
	✓		84.08	69.97	38.98	46.13 +1.12%
		✓	83.42	69.60	38.43	46.72 +2.41%
✓	✓	✓	**85.51**	**71.69**	**40.72**	47.67 +4.51%

(a) logic loss (§3.2)

# Iter.	mIoU^3↑ (multi scale)	mIoU^2↑ (multi scale)	mIoU^1↑ (multi scale)	Inference Speed (fps)
0	84.62	70.95	40.18	3.44
1	85.23	71.46	40.56	3.37 -2.03%
2	**85.51**	**71.69**	**40.72**	3.31 -3.78%
3	84.84	71.12	40.29	3.25 -5.52%
4	84.49	70.84	40.03	3.20 -6.98%

(b) iteration of message passing (§3.3)

q	mIoU^3↑ (multi scale)	mIoU^2↑ (multi scale)	mIoU^1↑ (multi scale)
1	83.83	70.22	38.77
3	84.65	71.15	40.09
5	**85.51**	**71.69**	**40.72**
8	84.47	71.03	39.74
10	83.52	69.88	38.25

(c) aggregation coefficient for $\forall$ (Eq.8)

Table 5: **Ablative studies** on Mapillary Vistas 2.0 [28] `val` (§4.4). All variants are based on DeepLabV3+ [32] with ResNet-101 [34] backbone.

based on DeepLabV3+ [32] with ResNet-101 [34] backbone.

Logic-Induced Training. We first study the effectiveness of our logic-induced training strategy (§3.2) in Table 5a. 1st row reports the results of our baseline model – DeepLabV3+. 2nd, 3rd, and 4th rows respectively list the scores obtained by individually adding our *C*-rule loss $\mathcal{L}_C$ (*cf.* Eq.10), *D*-rule loss $\mathcal{L}_D$ (*cf.* Eq.12), and *E*-rule loss $\mathcal{L}_E$ (*cf.* Eq.14). The last row gives the performance of our full loss $\mathcal{L}$ (*cf.* Eq.15). We can find that: **i)** Taking each of our logic losses into consideration can provide consistent performance gains. This demonstrates that different logic rules can describe different properties of semantic structure and verify that the segmentation model can indeed benefit from our proposed logic losses. **ii)** Combing all three logic losses together results in the best performance. This suggests that our logic rules provide a comprehensive description of the relational knowledge in the semantic hierarchy $\mathcal{T}$, and supports our core idea that exploiting symbolic knowledge is crucial for visual semantic interpretation and can boost sub-symbolic learning.

Training Speed. As shown in the last column of Table 5a, our logic-induced training regime causes a trivial delay (∼5.0%).

Logic-Induced Inference. We next investigate the impact of our logic-induced inference strategy (§3.3) in Table 5b. 1st row reports the results of network feed-forward output. The rest rows give the scores obtained with different iterations of message passing (*cf.* Eq. 17). These results demonstrate the efficacy of our strategy and the necessity of incorporating logic reasoning into network inference. We accordingly set 2-iteration as the default to pursue the best performance.

Inference Speed. We also report inference speed (fps) in Table 5b. As seen, our logic-induced inference strategy only slows the speed slightly during model deployment (∼3.8%).

Aggregation Coefficient. For the approximation of $\forall$ quantifier (*cf.* Eq. 8), we adopt the generalized mean for stable training, as suggested by [128]. Basically, a higher coefficient q renders $\forall$ a stronger focus on outliers. For completeness, the results with different values of q are reported in Table 5c.

5. Conclusion and Discussion

The creation of intelligent systems that integrate the fundamental cognitive abilities of reasoning and learning has long been viewed as a core challenge for AI [22]. While the community recently witnessed great advances in high-level perception tasks such as visual semantic interpretation, top-leading solutions are purely driven by sub-symbolic learning, far from such effective integration. The present study represents an innovative and solid attempt towards closing this gap. By embedding symbolic logic into both network training and inference, a structured and powerful visual semantic parser is delivered. We hope this work can stimulate our community to rethink current *de facto*, sub-symbolic paradigm and investigate new methodologies, from the perspective of achieving a better understanding of human and machine intelligence.

Acknowledgements This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938 and CCF-Tencent Open Fund.

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2, 7

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2

[3] Johannes Bill, Hrag Pailian, Samuel J Gershman, and Jan Drugowitsch. Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences*, 117(39):24581–24589, 2020. 1

[4] Steven M Frankland and Joshua D Greene. Concepts and compositionality: in search of the brain's language of thought. *Annual Review of Psychology*, 71:273–303, 2020. 1

[5] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, pages 1551–1558, 2021. 1

[6] Theo MV Janssen et al. Compositionality: Its historic context. *The Oxford handbook of compositionality*, pages 19–46, 2012. 1

[7] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1

[8] Peter Pagin and Dag Westerståhl. Compositionality i: Definitions and variants. *Philosophy Compass*, 5(3):250–264, 2010. 1

[9] Paul Smolensky, R Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. *Neurocompositional computing in human and machine intelligence: A tutorial*. Microsoft Technical Report MSR-TR-2022, 2022. 1

[10] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 2, 3, 6, 7

[11] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *CVPR*, 2018.

[12] Zhiheng Li, Wenxuan Bao, Jiayang Zheng, and Chenliang Xu. Deep grouping model for unified perceptual parsing. In *CVPR*, 2020. 3

[13] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *CVPR*, 2022. 2, 3, 4, 6, 7

[14] Wenguan Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *IEEE TPAMI*, 44(7):3508–3522, 2021. 2

[15] Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2:14–35, 2021. 2

[16] Stephen K Reed. A taxonomic analysis of abstraction. *Perspectives on Psychological Science*, 11(6):817–837, 2016. 2

[17] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011. 2

[18] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014. 2

[19] A Garcez, M Gori, LC Lamb, L Serafini, M Spranger, and SN Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4):611–632, 2019. 2, 3

[20] Henry Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Magazine*, 43(1):93–104, 2022.

[21] Wenguan Wang and Yi Yang. Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing. *arXiv preprint arXiv:2210.15889*, 2022. 2, 3

[22] Leslie G Valiant. Three problems in computer science. *Journal of the ACM*, 50(1):96–99, 2003. 2, 8

[23] Lazar Valkov, Dipak Chaudhari, Akash Srivastava, Charles Sutton, and Swarat Chaudhuri. Houdini: Lifelong learning as program synthesis. In *NeurIPS*, 2018. 2, 3

[24] Maxwell Nye, Armando Solar-Lezama, Josh Tenenbaum, and Brenden M Lake. Learning compositional rules via neural program synthesis. In *NeurIPS*, 2020.

[25] Emilio Parisotto, Abdel-rahman Mohamed, Rishabh Singh, Lihong Li, Dengyong Zhou, and Pushmeet Kohli. Neuro-symbolic program synthesis. In *ICLR*, 2017. 2, 3

[26] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *ICLR*, 2019. 2, 3

[27] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018. 2, 3

[28] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 6, 7, 8

[29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 6, 7

[30] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2, 6, 7

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2, 6, 7

[32] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 6, 7, 8

[33] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 6, 7, 8

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 6, 8

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 6

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[37] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[38] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 6, 7

[39] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Deep learning markov random field for semantic segmentation. *IEEE TPAMI*, 40(8):1814–1828, 2017.

[40] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

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

[42] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *CVPR*, 2019.

[43] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *ECCV*, 2020.

[44] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 6, 7

[45] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, 2020.

[46] Mingyuan Liu, Dan Schonfeld, and Wei Tang. Exploit visual dependency relations for semantic segmentation. In *CVPR*, 2021.

[47] Chi-Wei Hsiao, Cheng Sun, Hwann-Tzong Chen, and Min Sun. Specialize and fuse: Pyramidal output representation for semantic segmentation. In *ICCV*, 2021.

[48] Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *ICCV*, 2021.

[49] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *ICCV*, 2021.

[50] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 2020.

[51] Guangrui Li, Guoliang Kang, Xiaohan Wang, Yunchao Wei, and Yi Yang. Adversarially masking synthetic to mimic real: Adaptive noise injection for point cloud segmentation adaptation. In *CVPR*, 2023.

[52] Tianfei Zhou, Liulei Li, Xueyi Li, Chun-Mei Feng, Jianwu Li, and Ling Shao. Group-wise learning for weakly supervised semantic segmentation. *IEEE TIP*, 31:799–811, 2021. 2

[53] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016. 2

[54] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018.

[55] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *CVPR*, 2019.

[56] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *CVPR*, 2016.

[57] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *ECCV*, 2020.

[58] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, 2020. 2

[59] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *ICCV*, 2017. 2

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[61] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[62] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.

[63] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[64] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *ICCV*, 2019.

[65] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019.

[66] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 7

[67] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 6, 7

[68] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021.

[69] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021.

[70] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *NeurIPS*, 2022.

[71] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *ICCV*, 2023. 2

[72] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 2

[73] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc

Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022.
[74] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *NeurIPS*, 2022. 6, 7
[75] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. Boosting video object segmentation via space-time correspondence learning. In *CVPR*, 2023.
[76] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *CVPR*, 2023. 2
[77] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. Fasterseg: Searching for faster real-time semantic segmentation. *arXiv preprint arXiv:1912.10917*, 2019. 2
[78] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019.
[79] Vladimir Nekrasov, Hao Chen, Chunhua Shen, and Ian Reid. Fast neural architecture search of compact semantic segmentation models via auxiliary cells. In *CVPR*, 2019.
[80] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. Learning dynamic routing for semantic segmentation. In *CVPR*, 2020. 2
[81] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2, 6, 7
[82] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 7
[83] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 6, 7
[84] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 6, 7
[85] James Liang, Tianfei Zhou, Dongfang Liu, and Wenguan Wang. Clustseg: Clustering for universal segmentation. In *ICML*, 2023.
[86] Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, and Qinghua Zheng. Clustering based point cloud representation learning for 3d analysis. In *ICCV*, 2023. 2
[87] Terrence W Deacon. The co-evolution of language and the brain. *WW Norlon, Nueva Ymk*, 1997. 2
[88] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A modern approach*. Pearson, 3 edition, 2009. 2
[89] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019. 2, 3, 6
[90] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *CVPR*, 2020. 2, 3, 7
[91] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *CVPR*, 2022. 2, 3
[92] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2):113–140, 2005. 2
[93] Erik B Sudderth, Antonio Torralba, William T Freeman, and Alan S Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.
[94] Erik B Sudderth, Antonio Torralba, William T Freeman, and Alan S Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1-3):291–330, 2008.
[95] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing with attribute grammar. *IEEE TPAMI*, 31(1):59–73, 2008.
[96] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2
[97] Bingke Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Progressive cognitive human parsing. In *AAAI*, 2018. 3
[98] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. Learning semantic neural tree for human parsing. In *ECCV*, 2020. 3, 6
[99] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. 3
[100] Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1-2):77–105, 1990. 3
[101] Lokendra Shastri and Venkat Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16(3):417–451, 1993.
[102] Geoffrey G Towell and Jude W Shavlik. Knowledge-based artificial neural networks. *Artificial intelligence*, 70(1-2):119–165, 1994.
[103] Tony A Plate. Holographic reduced representations. *IEEE TNN*, 6(3):623–641, 1995. 3
[104] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 3
[105] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018. 3
[106] Eleonora Giunchiglia and Thomas Lukasiewicz. Multi-label classification neural networks with hard logical constraints. *Journal of Artificial Intelligence Research*, 72:759–818, 2021. 3
[107] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In *IJCAI*, 2022. 3
[108] Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *NeurIPS*, 2020. 3
[109] Komal Teru, Etienne Denis, and Will Hamilton. Inductive relation prediction by subgraph reasoning. In *ICML*, 2020.
[110] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP*, 2019. 3

[111] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *ACL*, 2016. 3

[112] Russell Stewart and Stefano Ermon. Label-free supervision of neural networks with physics and domain knowledge. In *AAAI*, 2017.

[113] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *ICML*, 2018. 3

[114] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *NeurIPS*, 2017. 3

[115] Michelangelo Diligenti, Marco Gori, and Claudio Sacca. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244:143–165, 2017.

[116] Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, and Martin Vechev. Dl2: Training and querying neural networks with logic. In *ICML*, 2019. 3

[117] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018. 3

[118] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In *NeurIPS*, 2019. 3

[119] Lili Mou, Zhengdong Lu, Hang Li, and Zhi Jin. Coupling distributed and symbolic execution for natural language queries. In *ICML*, 2017. 3

[120] Daoming Lyu, Fangkai Yang, Bo Liu, and Steven Gustafson. Sdrl: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *AAAI*, 2019.

[121] Fangkai Yang, Daoming Lyu, Bo Liu, and Steven Gustafson. Peorl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. In *IJCAI*, 2018. 3

[122] Forough Arabshahi, Sameer Singh, and Animashree Anandkumar. Combining symbolic expressions and black-box function evaluations in neural programs. In *ICLR*, 2018. 3

[123] Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Ying Nian Wu, and Song-Chun Zhu. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. In *ICML*, 2020.

[124] Guillaume Lample and François Charton. Deep learning for symbolic mathematics. In *ICLR*, 2019. 3

[125] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: neural probabilistic logic programming. In *NeurIPS*, 2018. 3

[126] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. In *NeurIPS*, 2019.

[127] Yaqi Xie, Ziwei Xu, Mohan S Kankanhalli, Kuldeep S Meel, and Harold Soh. Embedding symbolic knowledge into deep networks. In *NeurIPS*, 2019.

[128] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022. 3, 4, 5, 6, 8

[129] Wei Bi and James T Kwok. Multilabel classification on tree-and dag-structured hierarchies. In *ICML*, 2011. 4

[130] Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In *NeurIPS*, 2020. 4

[131] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *CVPR*, 2020. 4

[132] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *ICML*, 2018. 4

[133] Emile van Krieken, Erman Acar, and Frank van Harmelen. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302:103602, 2022. 4, 5

[134] Vilém Novák, Irina Perfilieva, and Jiri Mockor. *Mathematical principles of fuzzy logic*, volume 517. Springer Science & Business Media, 2012. 4

[135] Petr Hájek. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media, 2013. 4

[136] Solomon Feferman, John W Dawson Jr, Stephen C Kleene, Gregory H Moore, Robert M Solovay, and Jean Van Heijenoort. *Kurt Godel: collected works. Vol. 1: Publications 1929-1936*. Oxford University Press, Inc., 1986. 4

[137] Joseph A Goguen. The logic of inexact concepts. *Synthese*, pages 325–373, 1969. 5

[138] Umberto Michieli, Edoardo Borsato, Luca Rossi, and Pietro Zanuttigh. Gmnet: Graph matching network for large scale part semantic segmentation in the wild. In *ECCV*, 2020. 6, 7

[139] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Do semantic parts emerge in convolutional neural networks? *IJCV*, 126(5):476–494, 2018.

[140] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-class part parsing with joint boundary-semantic awareness. In *ICCV*, 2019. 7

[141] Rishubh Singh, Pranav Gupta, Pradeep Shenoy, and Ravikiran Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *CVPR*, 2022. 6, 7

[142] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 6

[143] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 6

[144] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014. 6

[145] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *CVPR*, 2019. 7

[146] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 7

[147] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021. 7