EasyChair Preprint

№ 4845

Commonsense Visual Sensemaking for Autonomous Driving: on Generalised Neurosymbolic Online Abduction Integrating Vision and Semantics

Jakob Suchan, Mehul Bhatt and Srikrishna Varadarajan

December 29, 2020

Commonsense Visual Sensemaking for Autonomous Driving

On Generalised Neurosymbolic Online Abduction Integrating Vision and Semantics

Jakob Suchan – Mehul Bhatt – Srikrishna Varadarajan

University of Bremen (Germany) — Örebro University (Sweden)

CoDesign Lab (EU) > Cognitive Vision
www.codesign-lab.org / info@codesign-lab.org

Abstract

We demonstrate the need and potential of systematically integrated *vision* and *semantics* solutions for visual sensemaking in the backdrop of autonomous driving. A general neurosymbolic method for *online* visual sensemaking using answer set programming (ASP) is systematically formalised and fully implemented. The method integrates state of the art in visual computing, and is developed as a modular framework that is generally usable within hybrid architectures for realtime perception and control. We evaluate and demonstrate with community established benchmarks KITTIMOD, MOT-2017, and MOT-2020. As use-case, we focus on the significance of human-centred visual sensemaking —e.g., involving semantic representation and explainability, question-answering, commonsense interpolation— in safety-critical autonomous driving situations. The developed neurosymbolic framework is domain-independent, with the case of autonomous driving designed to serve as an exemplar for online visual sensemaking in diverse cognitive interaction settings in the backdrop of select human-centred AI technology design considerations.

Keywords:
Cognitive Vision, Deep Semantics, Declarative Spatial Reasoning, Knowledge Representation and Reasoning, Commonsense Reasoning, Visual Abduction, Answer Set Programming, Autonomous Driving, Human-Centred Computing and Design, Standardisation in Driving Technology, Spatial Cognition and AI

PUBLICATION NOTE.
This is a preprint / review version of an accepted contribution to be published as part of the Artificial Intelligence Journal (AIJ).* The article is an extended version of an IJCAI 2019 publication [74]. The overall scientific agenda (pertaining to Cognitive Vision and Deep Semantics [12]) driving this research is available at:

CoDesign Lab (EU) > Cognitive Vision / https://codesign-lab.org/cognitive-vision/

Related select publications: https://codesign-lab.org/select-papers/#cognitive_vision

*The AIJ published version is final; it also fully incorporates all reviewer feedback.

1. MOTIVATION

Autonomous driving research has received enormous academic & industrial interest in recent years (Sec 5). This surge has coincided with (and been driven by) advances in *deep learning* based computer vision research. Although end-to-end deep learning based vision & control has (arguably) been successful for self-driving vehicles, we posit that there is a clear need and tremendous potential for hybrid visual sensemaking solutions that integrate *vision and semantics* towards fulfilling essential legal and ethical responsibilities involving explainability, human-centred AI (Artificial Intelligence), and industrial standardisation (e.g, pertaining to representation, realisation of rules and norms, fulfilling statutory obligations).

Autonomous Vehicles: "Standardisation and Regulation"

As the self-driving vehicle industry develops further, it will be necessary to have an articulation and community consensus on aspects such as representation, interoperability, human-centred performance benchmarks, and data archival & retrieval mechanisms. Within autonomous driving, the need for standardisation and ethical regulation has most recently garnered interest internationally, e.g., with the Federal Ministry of Transport and Digital Infrastructure in Germany (BMVI) taking a lead in eliciting 20 key propositions[1] (with legal implications) for the fulfilment of ethical commitments for automated and connected driving systems [20]. In spite of major investments in self-driving vehicle research, issues related to human-centred'ness, human collaboration, and standardisation have been barely addressed, with the current focus in driving research primarily being on two basic considerations: *how fast to drive, and which way and how much to steer*. This is necessary, but inadequate if autonomous vehicles are to become commonplace and function with humans [4, 22]. Ethically driven standardisation and regulation will require addressing challenges in foundational human-centred AI technology design, e.g., pertaining to semantic visual interpretation, natural / multimodal human-machine interaction, high-level data analytics (e.g., for post hoc diagnostics, dispute settlement). This will necessitate —amongst other things— human-centred qualitative benchmarks and design & evaluation of multifaceted hybrid solutions integrating diverse methodologies in Artificial Intelligence, Machine Learning, Cognitive Science, Design Science etc.

Neurosymbolism: Visual Sensemaking Needs Both "Vision and Semantics"

Visual sensemaking requires a systematically developed general and modular integration of high-level techniques concerned with "commonsense and semantics" with low-level neural methods capable of computing primitive features of interest in visual data. Towards this, this research demonstrates the significance of semantically-driven methods rooted in knowledge representation and reasoning (KR) in addressing research questions pertaining to explainability and human-centred AI particularly from the viewpoint of (perceptual) sensemaking of dynamic visual imagery. This is done in the backdrop of the autonomous driving domain; as an example, consider the *occlusion scenario* in Fig. 1:

> Car (c) is in-front, and indicating to turn-right; during this time, person (p) is on a bicycle (b) and positioned front-right of c and moving-forward. Car c turns-right, during which the bicyclist $< p, b >$ is not visible. Subsequently, bicyclist $< p, b >$ reappears.

[1] The 20 key propositions elicited by the German federal ministry BMVI highlight a range of factors pertaining to safety, utilitarian considerations, human rights, statutory liability, technological transparency, data management and privacy etc [20].

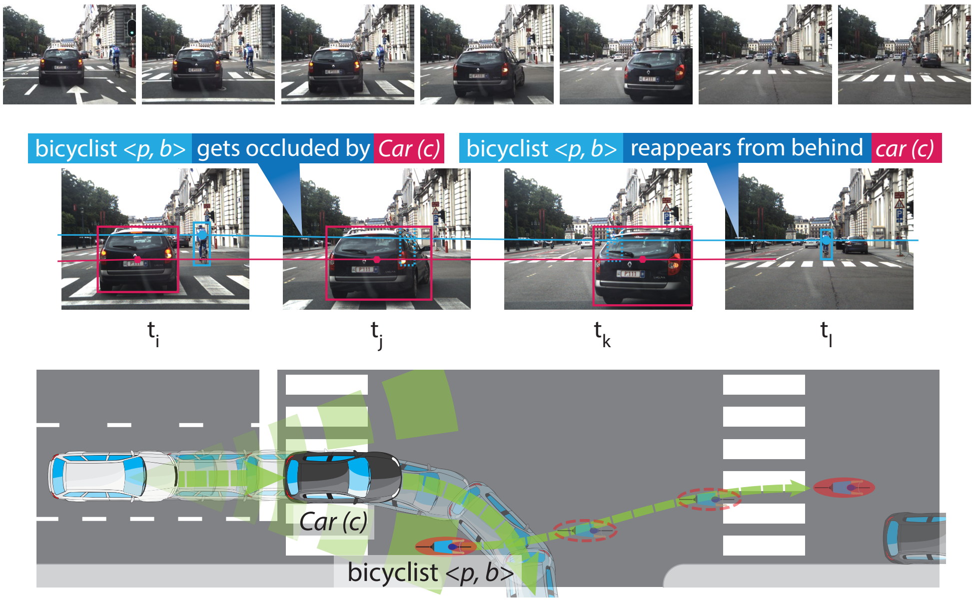


FIGURE 1: Out of sight but not out of mind; the case of hidden entities: e.g., an occluded cyclist.

The occlusion scenario of Fig. 1 is one of range of (seemingly) mundane safety-critical moments that one may regularly experience while driving a vehicle (Fig. 4, Fig. 7-8 and Table 6 include additional examples). This scenario is sufficiently indicative of several challenges concerning *epistemological* and *phenomenological* aspects relevant to a wide range of *dynamic spatial systems* [11, 14, 10]:

- **projection and interpolation** of missing information (e.g., what could be hypothesised about bicyclist $< p, b >$ when it is *occluded*; how can this hypothesis enable in planning
- object **identity maintenance** at a semantic level, e.g., in the presence of occlusions, missing and noisy quantitative data, error in detection and tracking
- ability to make **default assumptions**, e.g., pertaining to persistence objects and/or object attributes
- maintaining **consistent beliefs** respecting (domain-neutral) commonsense criteria, e.g., related to compositionality & indirect effects, space-time continuity, positional changes resulting from motion
- inferring / computing **counterfactuals**, in a manner akin to human cognitive ability to perform mental simulation for purposes of introspection, performing "what-if" reasoning tasks to determine an immediate next step).

Addressing such challenges —be it realtime or post-hoc— in view of human-centred AI concerns pertaining to representations rooted to natural language, explainability, ethics and regulation requires a systematic (neurosymbolic) integration of **Semantics and Vision**, i.e., robust commonsense representation & inference about spacetime dynamics on the one hand, and powerful low-level visual computing capabilities, e.g., pertaining to object detection and tracking on the other.

Deep Semantics: (Systematically) "Integrating AI and Vision"

The development of domain-independent computational models of perceptual sensemaking — e.g., encompassing capabilities such as visuospatial Q/A, spatio-temporal relational learning, visuospatial abduction— with multimodal human behavioural stimuli such as RGB(D), video, audio, eye-tracking requires the representational and inferential mediation of commonsense and spatio-linguistically rooted abstractions of space, motion, actions, events and interaction. We characterise **Deep Semantics** [12] within a declarative AI setting as:

- general methods for the processing and semantic interpretation of dynamic visuospatial imagery with an emphasis on the ability to **abstract, learn, and reason** with cognitively rooted structured characterisations of commonsense knowledge about **space and motion**.
- the existence of declarative models –e.g., pertaining to space, space-time, motion, actions & events, spatio-linguistic conceptual knowledge (e.g., Table 2)– and corresponding formalisation supporting (domain-neutral) **reasoning capabilities** (e.g., visual Q/A and learning, non-monotonic visuospatial abduction)

Formal semantics and computational models of deep semantics manifest themselves in declarative AI settings such as constraint logic programming, inductive logic programming, and answer set programming. Naturally, a practical illustration of the intergated "AI and Vision" method requires a tight but modular integration of the (declarative) commonsense spatio-temporal abstraction and reasoning with robust low-level visual computing foundations (primarily) driven by state of the art visual computing techniques (e.g., for visual feature detection, tracking).

KEY CONTRIBUTIONS

This research is situated within the broader auspices of the scientific agenda of cognitive vision and perception, which addresses visual, visuospatial and visuo-locomotive perception and interaction from the viewpoints of language, logic, spatial cognition and artificial intelligence [12] (Sec 5). The key contribution of this paper is to develop a general and systematic declarative visual sensemaking method capable of *online abduction*: **realtime, incremental, commonsense** question-answering and belief maintenance over dynamic visuospatial imagery. Supported are **(1–3)**:

(1). Human-Centred Representation for Space and Motion

Declaratively modelled ontological characterisation of human-centric relational representations that are semantically rooted to commonsense spatio-linguistic primitives pertaining to space and motion as they occur in natural language [16, 53].

(2). Systematic High-level Abductive Reasoning

Driven by Answer Set Programming (ASP) [23], the ability to abductively compute commonsense interpretations and explanations in a range of (a)typical everyday driving situations, e.g., concerning safety-critical decision-making; the declarative model of space and motion, in addition to supporting abductive reasoning about space and change, is also naturally amenable to high-level semantic interpretation (e.g., by question answering) for post-hoc analytical purposes (e.g., as might be relevant in situations requiring diagnosis et al for litigation, insurance claims).

(3). Online Performance of Modularly Integrated Vision and Semantics

Online performance –in an "active vision" context– of the overall framework modularly integrating high-level commonsense reasoning component with state of the art low-level (deep learning based) visual computing for practical application in real world settings (with autonomous driving serving as a solid demonstration platform).

ORGANISATION OF THE PAPER.

The rest of the article is organised as follows:

- **Section 2** presents the ontological and formal representational foundations of the developed visual sensemaking framework; main focus is on the commonsense representation aspects pertaining to the modelling of space, space-time, motion, events, and other aspects relevant to modelling and reasoning about spatio-temporal dynamics.
- **Section 3** presents the overall visual sensemaking framework and its technical implementation with a central focus on the general answer set programming based method for online abduction; we elaborate on the declarative model directly vis-a-vis the ASP implementation.
- **Section 4** demonstrates & empirically evaluates the core online abduction component with community established real-world datasets and benchmarks, namely: KITTIMOD [39], MOT-17 [54], and MOT-20 [31].
- **Section 5** discusses related works primarily from the viewpoints of knowledge representation, and visual computing as pursued in computer vision research.
- **Section 6** concludes with a brief summary of our work, together with pointers to immediate research questions for follow-up, as well as more broad-based directions that this work aims to open up.

Appendices A–C. AppendixA provides annotations of select Answer Set Programming source code relevant to the declarative model presented in Section 3. AppendixB presents additional examples chosen from community benchmark datasets together with sample data; it also includes an elaborated version of a running example used in the paper. AppendixC provide a succinct view of (select) data corresponding to (select) scenes.

2. COMMONSENSE – SPACE – MOTION: ONTOLOGICAL AND REPRESENTATIONAL ASPECTS

We present the ontological and formal representational foundations of the developed visual sensemaking framework while focussing on the commonsense representational aspects pertaining to the modelling of space, space-time, motion, events, and other aspects relevant to modelling and reasoning about spatio-temporal dynamics. Towards this, Table 1 summarises the individual constituents of Σ_{st} (spatiotemporal primitives) and Σ_{dyn} (spatiotemporal dynamics), and Table 2 elaborates the supported commonsense relations for the abstraction of space, motion, and (inter)action. Figure 3 is a (non-exhaustive) collection of generic / domain-neutral spacetime motion patterns supported; Figures 2 and 4 include concrete instance of such generic motion patterns: Fig. 2 illustrates motion patterns for *approach*, *occlusion*, and *connected motion*; and Fig. 4 illustrates the motion patterns underlying a security-critical scenario involved an elaborate lane changing episode.

2.1. Commonsense Abstractions for Space and Motion

Commonsense spatio-temporal relations and patterns (e.g., *left*, *touching*, *part of*, *during*, *collision*) offer a human-centered and cognitively adequate formalism for semantic grounding and automated reasoning for everyday (embodied) multimodal interactions [16, 53]. Qualitative, multi-domain[2] representations of spatial, temporal, and spatio-temporal relations and motion patterns (e.g., Fig 2-3), and their mutual transitions can provide a mapping between high-level semantic models of actions and events on one hand, and low-level / quantitative trajectory data emanating from visual computing algorithms on the other. For instance, by spatio-linguistically grounding complex trajectory data –e.g., pertaining to on-road moving objects– to a formal framework of space and motion, generalized (activity-based) commonsense reasoning about dynamic scenes, spatial relations, and motion trajectories denoting single and multi-object path & motion predicates can be supported. For instance, such predicates can be abstracted within a region-based 4D space-time framework [40, 6, 64], object interactions [27, 28], or even spatio-temporal narrative knowledge. An adequate commonsense spatio-temporal representation can, therefore, connect with low-level quantitative data, and also help to ground symbolic descriptions of actions and objects to be queried, reasoned about, or even manipulated in the real world.

2.2. Space, Motion, Objects, Events, Change: Ontology and Formal Model

Reasoning about spatio-temporal dynamics is based on high-level representations of objects, and their respective motion & mutual interactions in spacetime. Foundational ontological primitives for commonsense representation and reasoning about spatio-temporal dynamics are:

- Σ_{st} corresponds to primitives for representing space, time, motion and scene-level relational spatiotemporal structure
- Σ_{dyn} corresponds to the domain-independent commonsense theory for representing and reasoning about change.

$\mathbf{\Sigma_{st}}$ $<\mathcal{O}, \mathcal{E}, \mathcal{T}, \mathcal{MT}, \mathcal{R}>$ and $\mathbf{\Sigma_{dyn}}$ $<\Phi, \Theta>$ are as follows (Tables 1 and 2):

- **Domain Objects** ($\mathcal{O}$). The high-level, domain-dependent visual elements in the scene, e.g., road-side stakeholders such as *people*, *cars*, *cyclists*, constitute domain objects. Domain objects are denoted by $\mathcal{O} = \{o_1, ..., o_n\}$; elements in $\mathcal{O}$ are geometrically interpreted as *spatial entities*.

[2]Multi-domain refers to more than one aspect of space, e.g., topology, orientation, direction, distance, shape; this requires a mixed domain ontology involving points, line-segments, polygons, and regions of space, time, and space-time [80, 64, 40].

ONTOLOGY / SPACE & MOTION	REPRESENTATION	
Spatio-Temporal Ontology (Σ_{st})		
Domain Objects	$\mathcal{O} = \{o_1, ..., o_n\}$	e.g., *cars, people, cyclists*
Spatial Entities	$\mathcal{E} = \{\varepsilon_1, ..., \varepsilon_n\}$	*points, line-segments, rectangles*
Time	$\mathcal{T} = \{t_1, ..., t_n\}$	*time-points, time-intervals*
Motion	$\mathcal{MT}_{o_i} = (\varepsilon_{t_s}, ..., \varepsilon_{t_e})$	*motion tracks / space-time histories*
Spatio-Temporal Relationships	$\mathcal{R}$	*e.g., topology, orientation, distance*
Spatio-Temporal Dynamics (Σ_{dyn})		
Fluents	$\Phi = \{\phi_1, ..., \phi_n\}$	*e.g., visibility, hidden_by, clipped*
Events	$\Theta = \{\theta_1, ..., \theta_n\}$	*e.g., hides_behind, missing_detections*
Problem Specification		
Visual Observations	$\mathcal{VO}_t = \{obs_1, ..., obs_n\}$	*e.g., $\mathcal{E}$ corresponding to object detections*
Predictions	$\mathcal{P}_t = \{p_{trk_1}, ..., p_{trk_n}\}$	*e.g., $\mathcal{E}$ for predicted track*
Matching Likelihood	$\mathcal{ML}_t = \{ml_{trk_1,obs_1}, ..., ml_{trk_n,obs_m}\}$	*e.g., IoU between tracks and detections*
Hypothesis		
Assignments	$\mathcal{H}^{assign}$	*abduced assignments*
Events	$\mathcal{H}^{events} = \{\theta_1, ..., \theta_n\}$	*abduced event sequence*
Explanations	$\mathcal{EXP} \leftarrow < \mathcal{H}^{events}, \mathcal{MT} >$	*scene dynamics; abduced events and corresponding motion tracks*

TABLE 1: Commonsense – Space – Motion: Ontological and Representational Aspects

- **Spatial Entities** ($\mathcal{E}$). Spatial entities correspond to abstractions of domain objects by way of *points*, *line-segments* or (axis-aligned) *rectangles* based on their spatial properties (and a particular reasoning task at hand). Spatial entities are denoted by $\mathcal{E} = \{\varepsilon_1, ..., \varepsilon_n\}$.

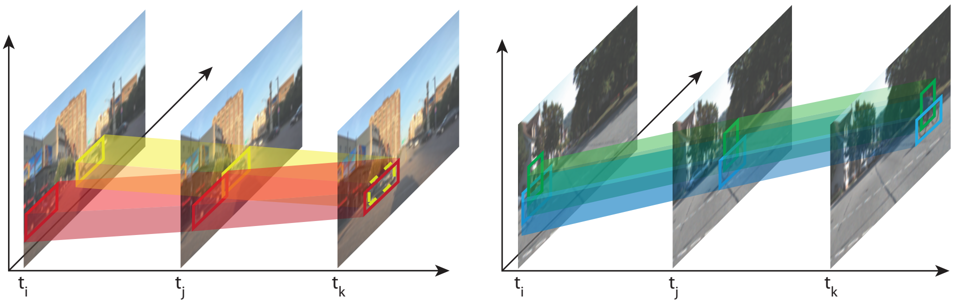
- **Time** ($\mathcal{T}$). The temporal dimension is represented by **time points**, denoted as $\mathcal{T} = \{t_1, ..., t_n\}$.

- **Motion Tracks** ($\mathcal{MT}$). Motion-tracks represent the spacetime motion trajectories (e.g., Fig. 2) of abstract spatial entities ($\mathcal{E}$) corresponding to domain object ($\mathcal{O}$) of interest. $\mathcal{MT}_{o_i} = (\varepsilon_{t_s}, ..., \varepsilon_{t_e})$ represents the **motion track** of a single object o_i, where t_s and t_e denote the start and end time of the track and ε_{t_s} to ε_{t_e} denotes the spatial entity ($\mathcal{E}$) —e.g., the *axis-aligned bounding box*—corresponding to the object o_i at time points t_s to t_e. Whereas Figures 2 and 4 presents one example of a space-time trajectory, Fig. 3 is a general (but non-exhaustive) set of patterns supported by our reasoning framework.

- **Spatio-Temporal Relationships** ($\mathcal{R}$). The spatial configuration of the scene and changes thereof are characterised based on the spatio-temporal relationships ($\mathcal{R}$; Table 2) between abstract representations ($\mathcal{E}$) of the domain objects ($\mathcal{O}$). For the running and demo examples of this paper, positional relations on axis-aligned rectangles based on the Rectangle Algebra (RA) [5] suffice; RA uses the relations of Interval Algebra (IA) [2] $\mathcal{R}_{IA} \equiv$ {before,

SPATIO-TEMPORAL DOMAIN ($\mathcal{QS}$)	Spatial, Time, Motion Relations ($\mathcal{R}$)	Entities ($\mathcal{E}$)
Mereotopology	disconnected (dc), external contact (ec), partial overlap (po), tangential proper part (tpp), non-tangential proper part (ntpp), proper part (pp), part of (p), discrete (dr), overlap (o), contact (c)	arbitrary rectangles, circles, polygons, cuboids, spheres
Incidence	interior, on boundary, exterior, discrete, intersects	2D point with rectangles, circles, polygons; 3D point with cuboids, spheres
Orientation	left, right, collinear, front, back, on, facing towards, facing away, same direction, opposite direction	2D point, circle, polygon with 2D line
Distance, Size	adjacent, near, far, smaller, equi-sized, larger	rectangles, circles, polygons, cuboids, spheres
Motion	moving: towards, away, parallel; growing / shrinking: vertically, horizontally; splitting / merging; rotation: left, right, up, down, clockwise, couter-clockwise	rectangles, circles, polygons, cuboids, spheres
Time	before, after, meets, overlaps, starts, during, finishes, equals	time-points, time intervals

TABLE 2: Commonsense Relations for Abstract Representation of Space, Motion, Interaction



a) Moving Towards and Occluding (Two cars crossing eachother)

b) Connected Motion (Person on a Bicycle)

FIGURE 2: Space-Time Histories in Context: Motion Tracks Under Conditions of Occlusion and Partial Overlapp

after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal} to relate two objects by the *interval relations* projected along each modelled dimension separately (e.g., horizontal and vertical dimensions).

- **Dynamics / Fluents and Events.** The set of **fluents** $\Phi = \{\phi_1, ..., \phi_n\}$ and **events** $\Theta = \{\theta_1, ..., \theta_n\}$ respectively characterise the dynamic properties of the objects in the scene and high-level abducibles (e.g., Tables 4 and 5). For reasoning about dynamics (with <Φ, Θ>), we use the epistemic generalisation of the event calculus [47] as per the formalisation in [51, 55]; in particular, for examples of this paper, the Functional Event Calculus (FEC) fragment of Ma et al. [51] suffices.[3]

[3] Main axioms relevant for this paper pertain to occurs-at(θ, t) denoting that an event occurred at time t and holds-at(ϕ, v, t) denoting that v holds for a fluent ϕ at time t. It it worth noting that in so far as the approach to reason about changes is concerned, our modular framework is by no means limited to the specific approach being utilised.

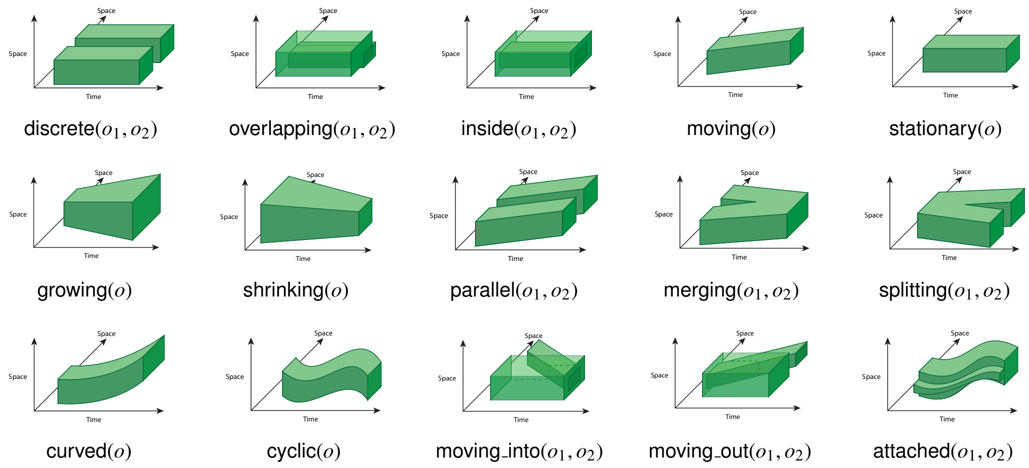


FIGURE 3: Commonsense Spatial Reasoning with Spatio-Temporal Entities. Illustrated are: Space-Time Histories for Spatio-temporal Patterns and Events

Problem Specification and Hypothesis.

- **Problem Specification** $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$. The abduction for each time point is given by the visual observations ($\mathcal{VO}_t$) consisting of spatial entities $\mathcal{E}$, i.e., bounding boxes for the detected objects, spatial entities $\mathcal{E}$ of object detections; the predicted locations ($\mathcal{P}_t$) for each track at time point t given as spatial entities $\mathcal{E}$; and the matching likelihood ($\mathcal{ML}_t$), i.e., based on the IoU between detected objects and tracks, providing an estimate of how likely a detection belongs to a track,.

- **Hypothesis** Abduced hypothesis consist of assignments ($\mathcal{H}^{assign}$) of detections to tracks and high-level events ($\mathcal{H}^{events}$) explaining object motion, e.g., occlusion of an object, caused by the object passing behind an other object. The online abduction results in abduced visuo-spatial dynamics ($\mathcal{EXP}$) consisting of motion tracks ($\mathcal{MT}$) (generated using the abduced assignments in $\mathcal{H}^{assign}$) and the events ($\mathcal{H}^{events}$) explaining the motion tracks.

3. VISUAL SENSEMAKING: A GENERAL METHOD DRIVEN BY ANSWER SET PROGRAMMING

Rooted in answer set programming, the developed framework is general, modular, and designed for integration as a reasoning engine within (hybrid) architectures designed for real-time decision-making and control where visual perception is needed as one of the several components. In such large scale AI systems the declarative model of the scene dynamics resulting from the presented framework can be used for semantic question-answering (Q/A), inference etc to support decision-making.

In principle, any method capable of modelling *dynamic spatial systems* [11] encompassing *space, actions, and change*

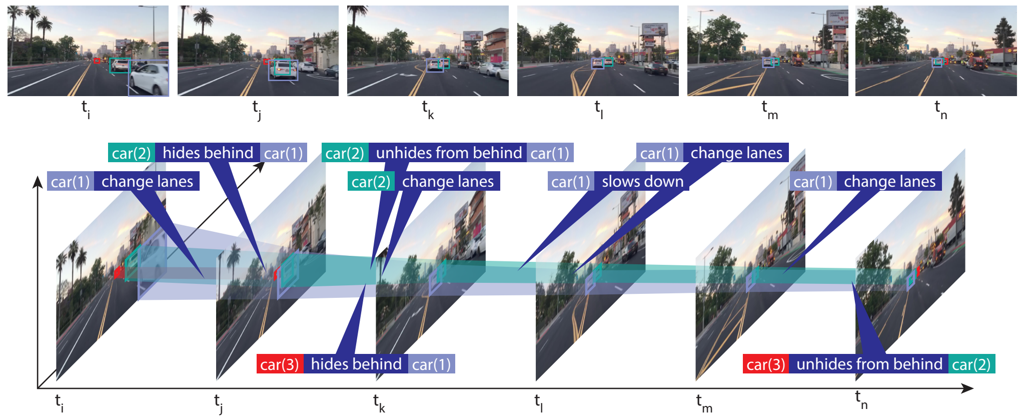


FIGURE 4: Space-Time Histories of Moving Objects: Safety-Criticality Case of a Close Encounter / Car(1) is moving towards car(2) on the right lane, and changes to the left lane to perform an overtaking action; subsequently, car(2) also changes to left lane to overtake car(3) that stopped and is blocking the right lane. To avoid a collision car(1) performs an emergency break and leaves the left lane to the left, entering the lane for the oncoming traffic.

3.1. Tracking as Abduction

Our proposed framework, in essence, jointly solves the problem of assignment of *detections* to *tracks* and explaining overall scene dynamics (e.g. appearance, disappearance) in terms of high-level *events* within an online integrated low-level visual computing and high-level abductive reasoning framework (Fig. 5).

Scene dynamics are tracked using a *detect and track* approach: we tightly integrate low-level visual computing (for detecting scene elements) with high-level ASP-based abduction to solve the assignment of observations to object tracks in an *incremental* manner. For each time point t we generate a *problem specification* consisting of the object tracks and visual observations and use ASP to abductively solve the corresponding assignment problem incorporating the ontological structure of the domain / data (abstracted with Σ).

Steps 1–3 (Alg. 1 & Table 3) consist of:[4]

1) Formulating the ASP problem specification consisting of the visual observations, prediction of motion of each object, and a measure for the likelyhood that a detection is associated with a track. Further the problem specification contains the state of the world, given by the sequence of events ($\mathcal{H}^{events}$) before time point t.

2) Associating detections to tracks, by jointly abducing matchings between object detections and tracks, together with the high-level events explaining these matches.

[10, 14] is usable; basic considerations guiding choice of an action theory pertain to expressivity, modular elaboration tolerance, and support for basic epistemological aspects such as *frame* and *ramification* [65]. For instance, other epistemic settings for abductive inference with ASP too may be utilised [33, 34].

[4] In the context of Alg. 1 / Table 3, note that we utilise Clingo v5.3.0 [38] for the grounding and solving of the answer set program.

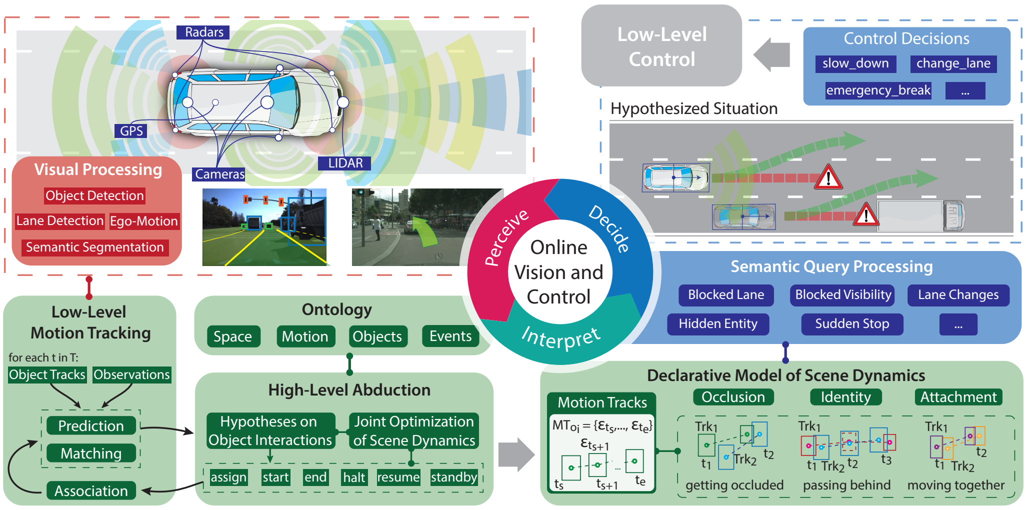


FIGURE 5: A General Online Abduction Framework / Conceptual Overview

3) Finding the hypothesis and corresponding associations best explaining the visual observations using optimization, i.e., maximizing matching likelihood and minimizing event costs.

In the following we describe each step in detail:

Step 1. Formulating the Problem Specification

The ASP problem specification for each time point t is given by the tuple $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$ and the sequence of events ($\mathcal{H}^{events}$) before time point t.

• **Visual Observations** Scene elements derived directly from the visual input data are represented as spatial entities $\mathcal{E}$, i.e., $\mathcal{VO}_t = \{\varepsilon_{obs_1}, ..., \varepsilon_{obs_n}\}$ is the set of observations at time t (Table 3). For the examples and empirical evaluation in this paper (Sec. 4) we focus on *Obstacle / Object Detections* – detecting cars, pedestrians, cyclists, traffic lights etc using YOLOv3 [60]. Further we generate scene context using *Semantic Segmentation* – segmenting the road, sidewalk, buildings, cars, people, trees, etc. using DeepLabv3+ [24], and *Lane Detection* – estimating lane markings, to detect lanes on the road, using SCNN [57]. Type and confidence score for each observation is given by $type_{obs_i}$ and $conf_{obs_i}$.

• **Movement Prediction** For each track trk_i changes in *position* and *size* are predicted using kalman filters; this results in an estimate of the spatial entity ε for the next time-point t of each motion track $\mathcal{P}_t = \{\varepsilon_{trk_1}, ..., \varepsilon_{trk_n}\}$.

• **Matching Likelihood** For each pair of tracks and observations ε_{trk_i} and ε_{obs_j}, where $\varepsilon_{trk_i} \in \mathcal{P}_t$ and $\varepsilon_{obs_j} \in \mathcal{VO}_t$, we compute the likelihood $\mathcal{ML}_t = \{ml_{trk_1,obs_1}, ..., ml_{trk_i,obs_j}\}$ that ε_{obs_j} belongs to ε_{trk_i}. The intersection over union (IoU) provides a measure for the amount of overlap between the *spatial entities* ε_{obs_j} and ε_{trk_i}.

Step 2. Abduction based Association Following perception as logical abduction most directly in the sense of Shanahan [66], we define the task of abducing visual explanations as find-

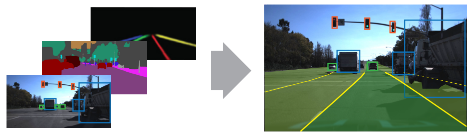
For each $t \in T$

Step 1. **Formulating the Problem Specification** $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$

(1) detect Visual Observations ($\mathcal{VO}_t$) e.g., *People, Cars, Objects, Roads, Lanes*,

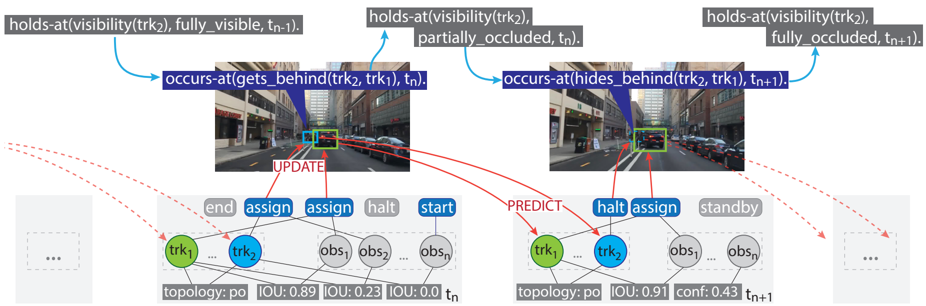
(2) Predictions ($\mathcal{P}_t$) of next position and size of object tracks using kalman filters, and

(3) calculate Matching Likelihood ($\mathcal{ML}_t$) based on Intersection over Union (IoU) between predictions and detections.



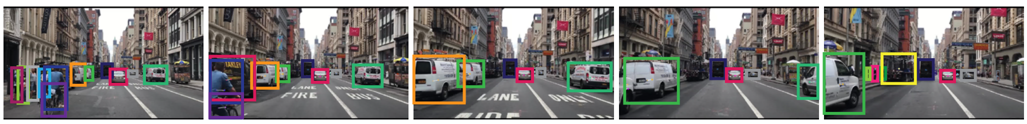
```
obs(obs_0,car,99). obs(...). ... box2d(obs_16,1078,86,30,44). ...
trk(trk_0,car). trk(...). ... box2d(trk_0,798,146,113,203). ...
iou(trk_0,obs_0,83921). iou(...). ... iou(trk_23,obs_16,0). ...
```

Step 2. **Abduction based Association** generate hypothesis for **(1)** matching of tracks and observations ($\mathcal{H}_t^{assign}$), and **(2)** and high-level events ($\mathcal{H}_t^{events}$) explaining **(1)**.



Step 3. **Finding the Optimal Hypothesis** Jointly optimize $\mathcal{H}_t^{assign}$ and $\mathcal{H}_t^{events}$ by maximizing matching likelihood $\mathcal{ML}_t$ and minimizing event costs.

RESULT. **Visuo-Spatial Scene Semantics** Resulting motion tracks ($\mathcal{MT}$) and the corresponding event sequence ($\mathcal{H}^{events}$) explaining the low-level motion:



```
... occurs_at(missing_detections(trk_10),35) ...  occurs_at(recover(trk_10),36)
... occurs_at(lost(trk_18),41) ... occurs_at(hides_behind(trk_9,trk_10),41)
... occurs_at(...) ... occurs_at(unhides_from_behind(trk_9,trk_10),42) ...
```

TABLE 3: Computational Steps for Online Visual Abduction

Algorithm 1: Online_Abduction($\mathcal{V}$, Σ)

Data: Visual imagery ($\mathcal{V}$), and
background knowledge $\Sigma \equiv_{def} \Sigma_{dyn} \cup \Sigma_{st}$

Result: Visual Explanations ($\mathcal{EXP}$) (Refer Table 3)

$\mathcal{MT} \leftarrow \varnothing, \mathcal{H}^{events} \leftarrow \varnothing$

for $t \in T$ **do**

  $\mathcal{VO}_t \leftarrow observe(\mathcal{V}_t)$

  $\mathcal{P}_t \leftarrow \varnothing, \mathcal{ML}_t \leftarrow \varnothing$

  for $trk \in \mathcal{MT}_{t-1}$ **do**

    $p_{trk} \leftarrow kalman\_predict(trk)$

    $\mathcal{P}_t \leftarrow \mathcal{P}_t \cup p_{trk}$

    for $obs \in \mathcal{VO}_t$ **do**

      $ml_{trk,obs} \leftarrow calc\_IoU(p_{trk}, obs)$

      $\mathcal{ML}_t \leftarrow \mathcal{ML}_t \cup ml_{trk,obs}$

  $abduce(< \mathcal{H}_t^{assign}, \mathcal{H}_t^{events} >)$ (Refer Step 2)

  $\mathcal{H}^{events} \leftarrow \mathcal{H}^{events} \cup \mathcal{H}_t^{events}$

  $\mathcal{MT}_t \leftarrow update(\mathcal{MT}_{t-1}, \mathcal{VO}_t, \mathcal{H}_{assign})$

return $\mathcal{EXP} \leftarrow < \mathcal{H}^{events}, \mathcal{MT} >$

ing an association ($\mathcal{H}_t^{assign}$) of observed scene elements ($\mathcal{VO}_t$) to the motion tracks of objects ($\mathcal{MT}$) given by the predictions $\mathcal{P}_t$, together with a high-level explanation ($\mathcal{H}_t^{events}$), such that $[\mathcal{H}_t^{assign} \wedge \mathcal{H}_t^{events}]$ is consistent with the background knowledge and the previously abduced event sequence $\mathcal{H}^{events}$, and entails the perceived scene given by $< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$:

$$\Sigma \wedge \mathcal{H}^{events} \wedge [\mathcal{H}_t^{assign} \wedge \mathcal{H}_t^{events}] \models \mathcal{VO}_t \wedge \mathcal{P}_t \wedge \mathcal{ML}_t$$

where $\mathcal{H}_t^{assign}$ consists of the assignment of detections to object tracks, and $\mathcal{H}_t^{events}$ consists of the high-level *events* Θ explaining the assignments.

• **Associating Objects and Observations** Finding the best match between observations ($\mathcal{VO}_t$) and object tracks ($\mathcal{P}_t$) is done by generating all possible assignments and then maximising a matching likelihood ml_{trk_i,obs_j} between pairs of spatial entities for matched observations ε_{obs_j} and predicted track region ε_{trk_i} (See Step 3). Towards this we use *choice rules* [38] (i.e., one of the heads of the rule has to be in the stable model) for ε_{obs_j} and ε_{trk_i}, generating all possible assignments in terms of assignment actions: *assign, start, end, halt, resume, ignore_det, ignore_trk*.

```
{
    assign(Trk, Det): det(Det, _, _);
    end(Trk);
    ignore_trk(Trk);
    halt(Trk);
    resume(Trk, Det): det(Det, _, _)
```

EVENTS	Description
enters_fov(Trk)	Track Trk enters the field of view.
leaves_fov(Trk)	Track Trk leaves the field of view.
hides_behind(Trk_1, Trk_2)	Track Trk_1 hides behind track Trk_2.
unhides_from_behind(Trk_1, Trk_2)	Track Trk_1 unhides from behind track Trk_2.
missing_detections(Trk)	Missing detections for track Trk.

TABLE 4: ABDUCIBLES; Events Relevant to Explaining (Dis)Appearance

FLUENTS	Values	Description
in_fov(Trk)	{true;false}	Track Trk is in the field of view.
hidden_by(Trk1, Trk2)	{true;false}	Track Trk1 is hidden by Trk2.
visibility(Trk)	{fully_visible; partially_occluded; fully_occluded}	Visibility state of track Trk.
clipped(Trk)	{true;false}	Track Trk is interrupted, e.g., missing detection(s).

TABLE 5: ABDUCIBLES; Fluents Relevant to Explaining (Dis)Appearance

```
}1
:- trk(Trk, _).
```

```
1{
    assign(Trk, Det): trk(Trk, _);
    start(Det);
    ignore_det(Det);
    resume(Trk, Det): trk(Trk, _)
}1
:- det(Det, _, _).
```

For each assignment action we define *integrity constraints*[5] that restrict the set of answers generated by the choice rules, e.g., the following constraints are applied to assigning an observation ε_{obs_j} to a track trk_i, applying thresholds on the IoU_{trk_i,obs_j} and the confidence of the observation $conf_{obs_j}$, further we define that the type of the observation has to match the type of the track it is assigned to (e.g., also see Fig. 6):

```
:- assign(Trk, Det), not assignment_constraints(Trk, Det).
```

```
assignment_constraints(Trk, Det) :-
    trk(Trk, Trk_Type), trk_state(Trk, active),
    det(Det, Det_Type, Conf), Conf > conf_thresh_assign,
    match_type(Trk_Type, Det_Type),
    iou(Trk, Det, IOU), IOU > iou_thresh.
```

[5] Integrity constraints restrict the set of answers by eliminating stable models where the body is satisfied.

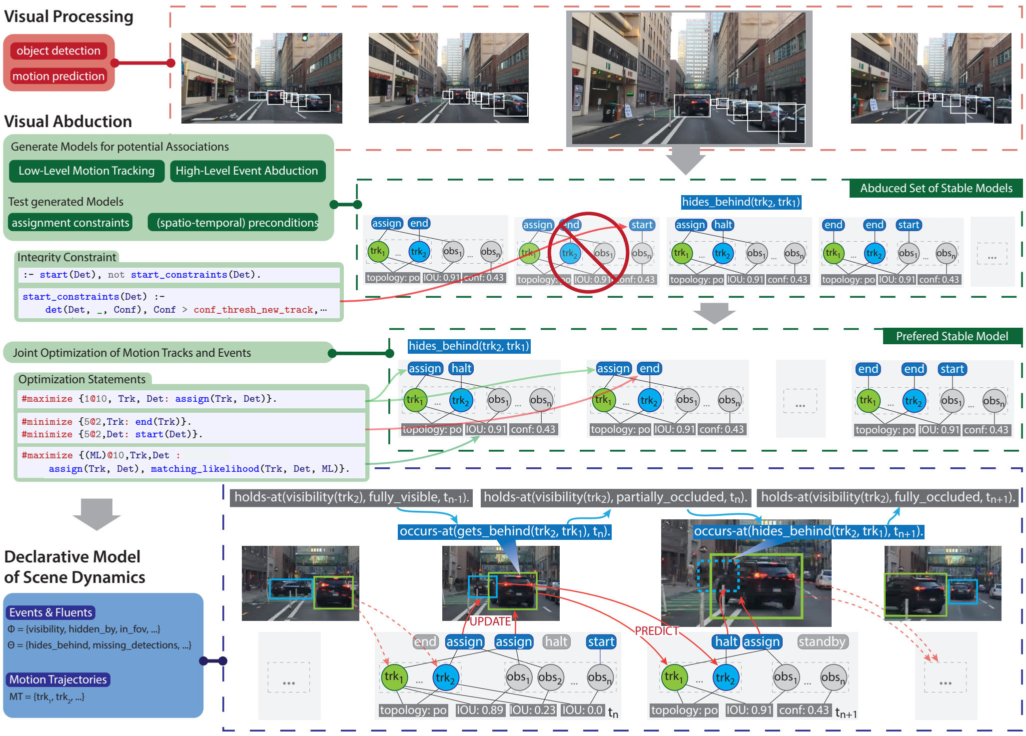


FIGURE 6: Commonsense Visual Explainability in Active Vision & Control (for Autonomous Driving); The Case of Hidden Entities.

• **Abducible High-Level Events** For the length of this paper, we restrict to high-level visuo-spatial abducibles pertaining to *object persistence* and *visibility* (Table 4): **(1)**. *Occlusion*: Objects can disappear or reappear as result of occlusion with other objects; **(2)**. *Noise and Missing Observation*: (Missing-)observations can be the result of faulty detections.

Lets take the case of *occlusion*: functional fluent visibility could be denoted *fully_visible*, *partially_occluded* or *fully_occluded*:

```
fluent(visibility(Trk)) :- trk(Trk, _).
```

```
possVal(visibility(Trk), fully_visible) :- trk(Trk, _).
possVal(visibility(Trk), partially_visible) :- trk(Trk,_).
possVal(visibility(Trk), not_visible) :- trk(Trk, _).
```

We define the event *hides_behind*/2, stating that an object hides behind another object by defining the conditions that have to hold for the event to possibly occur, and the effects the occurrence of the event has on the properties of the objects, i.e., the value of the visibility fluent changes to *fully_occluded*.

```
event(hides_behind(Trk1,Trk2)) :- trk(Trk1,_),trk(Trk2,_).
```

```
causesValue(hides_behind(Trk1, Trk2), visibility(Trk1), not_visible, T) :-
    trk(Trk1,_), trk(Trk2,_), time(T).
```

```
:- occurs_at(hides_behind(Trk1, Trk2), curr_time), not poss(hides_behind(Trk1, Trk2)).

poss(hides_behind(Trk1, Trk2)) :-
    trk(Trk1, _), trk(Trk2, _),
    position(overlapping_top, Trk1, Trk2),
    not holds_at(visibility(Trk1), not_visible, curr_time),
    not holds_at(visibility(Trk2), not_visible, curr_time).
```

For abducing the occurrence of an event we use choice rules that connect the event with assignment actions, e.g., a track getting halted may be explained by the event that the track hides behind another track.

```
1{
    occurs_at(hides_behind(Trk, Trk2), curr_time):  trk(Trk2,_);
    ...
}1
:-   halt(Trk).
```

Step 3. **Finding the Optimal Hypothesis** To ensure an *optimal assignment*, we use ASP based optimization to maximize the matching likelihood between matched pairs of tracks and detections. Towards this, we first define the matching likelihood based on the Intersection over Union (IoU) between the observations and the predicted boxes for each track as described in [9]:

```
matching_likelihood(Trk, Det, IOU) :-
    det(Det, _, _), trk(Trk, _), iou(Trk, Det, IOU).
```

We then maximize the matching likelihood for all assignments, using the build in *maximize* statement:

```
#maximize {(ML)@10,Trk,Det : assign(Trk, Det), matching_likelihood(Trk, Det, ML)}.
```

To find the best set of hypotheses with respect to the observations, we *minimize* the occurrence of certain events and association actions, e.g., the following optimization statements minimize starting and ending tracks; the resulting assignment is then used to update the motion tracks accordingly.

```
#minimize {5@2,Trk: end(Trk)}.
#minimize {5@2,Det: start(Det)}.
```

It is important here to note that: **(1)**. by jointly abducing the object dynamics and high-level events we can impose constraints on the assignment of detections to tracks, i.e., an assignment is only possible if we can find an explanation supporting the assignment; and **(2)**. the likelihood that an event occurs guides the assignments of observations to tracks. Instead of independently tracking objects and interpreting the interactions, this yields to event sequences that are consistent with the abduced object tracks, and noise in the observations is reduced (See evaluation in Sec. 4).

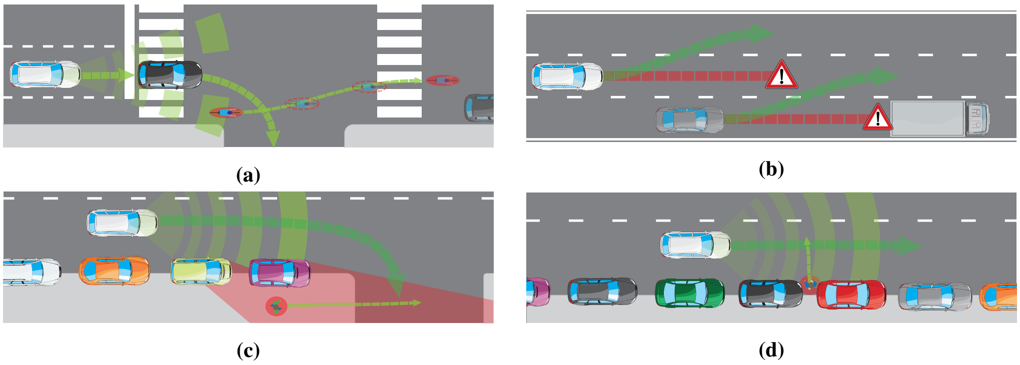


FIGURE 7: Safety-Critical Situation (select prototypes): **(a)**. momentarily occluded / hidden entities; **(b)**. overtaking / lane-crossing situation; **(c)**. blocked visibility; and **(d)** suddenly appearing objects.

SITUATION	Objects	Description
OVERTAKING	*vehicle, vehicle*	vehicle is overtaking another vehicle in front of the car.
HIDDEN_ENTITY	*entity, object*	traffic participant may be hidden by an obstacle, e.g. another car or van.
REDUCED_VISIBILITY	*object*	visibility is reduced by some object in front of car.
SUDDEN_STOP	*vehical*	vehicle in front of the car is suddenly stopping.
BLOCKED_LANE	*lane, object*	lane of the road is blocked by some object.
EXITING_PARKED_VEHICLE	*person, vehicle*	person is exiting a parked vehicle.

TABLE 6: Select Safety-Critical Situations

4. EVALUATION: APPLICATION AND EMPIRICAL PERFORMANCE ANALYSIS

We demonstrate applicability towards identifying and interpreting *safety-critical situations* (e.g., Table 6; Figures 7, 8; Fig. 4); these encompass those scenarios where interpretation of spacetime dynamics, driving behaviour, environmental characteristics is necessary to anticipate and avoid potential dangers. We also provide an empirical evaluation of the active sensemaking framework in the context of community benchmark datasets.

4.1. Application: Visual Perception by Abduction

4.1.1. Abducing Explanations: Appearance and Disappearance

Consider the scene in Fig. 9, where a car is passing behind a bus and is getting hidden during this. When the car hides behind the bus (time point 235), the track *trk_13* gets *halted* and the event *hides_behind* is abduced to explain why the car is not detected anymore and the corresponding track is halted.

The problem specification for time point 235 ($< \mathcal{VO}_{235}, \mathcal{P}_{235}, \mathcal{ML}_{235} >$) is given as follows:

(a)

(b)

(c)

(d)

(e)

FIGURE 8: Sample Safety-Critical Episodes: **(a)**. overtaking event in front of the car; **(b)**. occlusion while turning left; **(c)**. abrupt lane change on the highway; **(d)**. pedestrian suddenly appearing from between two parked cars; and **(e)**. (relatively) crowded and chaotic inner city traffic.

- $\mathcal{VO}_{235}$ the visual observation, consisting of the object detections, given by the bounding box, the type and the confidence:

```
det(det_0, person, 99). det(det_1, bus, 99). det(det_2, traffic_light, 86).
det(det_3, traffic_light, 81). det(det_4, traffic_light, 78).
det(det_5, traffic_light, 59).

box2d(det_0, 1114, 450, 148, 270). box2d(det_1, 8, 305, 992, 333).
box2d(det_2, 656, 205, 21, 56). box2d(det_3, 179, 137, 42, 75).
box2d(det_4, 108, 89, 46, 86). box2d(det_5, 784, 202, 21, 44).
```

- $\mathcal{P}_{235}$ the predictions for each track, given by the predicted bounding box, the state in which the track currently is, and the type of the tracked object:

```
trk(trk_3, traffic_light). trk_state(trk_3, active). trk(trk_7, traffic_light).
trk_state(trk_7, active). trk(trk_8, traffic_light). trk_state(trk_8, active).
trk(trk_12, bus). trk_state(trk_12, active). trk(trk_13, car).
trk_state(trk_13, active). trk(trk_15, person). trk_state(trk_15, active).

box2d(trk_3, 178, 136, 43, 73). box2d(trk_7, 105, 90, 49, 82).
box2d(trk_8, 655, 205, 21, 55). box2d(trk_12, 48, 294, 915, 350).
```

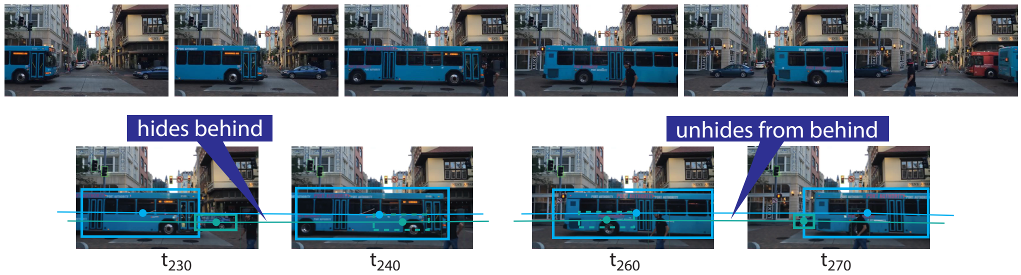


FIGURE 9: Abducing 'Hiding Behind' Event

```
box2d(trk_13, 904, 473, 181, 108). box2d(trk_15, 1111, 427, 156, 310).
```

- And $\mathcal{ML}_{235}$ the matching likelihood for each track with each detection, here given by the IoU between the detection bounding box and the predicted bounding box for the track[6]:

```
iou(trk_15,det_0,82426). iou(trk_12,det_1,88079). iou(trk_13,det_1,3022).
iou(trk_8,det_2,98532). iou(trk_3,det_3,94981). iou(trk_7,det_4,90457).
```

Solving the assignment of detections to tracks can now be done based on the *choice rules* for associating objects and observations, detailed in Section 3.1 Step 2.

To restrict the assignment we can impose constraints on the matching, by stating integrity constraints, e.g., for ensuring that only tracks and detections with the same type are matched, we could state the following integrity constraint. Stating that any stable model where the body is satisfied can not be in the set of answers, i.e., any model assigning a track and a detection which are not of the same type can not be an answer. Further, the track has to be active, the confidence of the detection has to be above a threshold, and the IoU between the track and the detection has to be above a threshold:

```
:- assign(Trk, Det), not assignment_constraints(Trk, Det).

assignment_constraints(Trk, Det) :-
    trk(Trk, Trk_Type), det(Det, Det_Type, Conf),
    trk_state(Trk, active),
    match_type(Trk_Type, Det_Type),
    Conf > conf_thresh_assign,
    iou(Trk, Det, IOU), IOU > iou_thresh.
```

By maximizing the matching likelihood we get the optimal assignment of detections to tracks, in our example the bus is detected by detection *det_1* which gets assigned to the corresponding track *trk_12*, but as the car is hiding behind the bus, there is no corresponding detection, thus the track of the car *trk_13* gets halted:

[6] Note, only those IoUs are stated which are bigger than 0.

```
halt(trk_13) assign(trk_15,det_0) assign(trk_12,det_1) assign(trk_8,det_2)
assign(trk_3,det_3) assign(trk_7,det_4)
```

The assignment actions are linked with high-level events for explaining the assignments, i.e., the halted track *trk_13* can be explained either by missing detections or by the track hiding behind another track. In this case track *trk_13* is hiding behind track *trk_12*, this can be abduced based on possible events, which in this case is the *hides_behind* event.

For the event *hides_behind*\2 the predicted tracks have to be overlapping. This is ensured by (spatial) preconditions of the event, given by the predicate *poss*\1:

```
poss(hides_behind(Trk1, Trk2)) :-
    trk(Trk1, _), trk(Trk2, _),
    position(overlapping_top, Trk1, Trk2),
    not holds_at(visibility(Trk1), not_visible, curr_time),
    not holds_at(visibility(Trk2), not_visible, curr_time).
```

In our example we can now abduce that the track *trk_13* representing the car is ended, because the car got hidden by the bus represented by track *trk_12*. In the formal representation of event calculus this is represented by the predicate *occurs_at*\2 as follows:

```
occurs_at(hides_behind(trk_13,trk_12),235)
```

At time point 268 the car reappears, after passing behind the bus. Due to the previously abduced event *hides_behind*\2, the visibility fluent for the track of the car *trk_13* has now the value *not_visible*.

For the detection *det_1* we can then abduce that track *trk_13* unhides from behind track *trk_12* based on the following event definition, stating that the event *unhides_from_behind*\2 is possible when Trk1 is *not_visible* and Trk2 is not *not_visible*:

```
poss(unhides_from_behind(Trk1, Trk2)) :-
    trk(Trk1, _), trk(Trk2, _),
    holds_at(visibility(Trk1), not_visible, curr_time),
    not holds_at(visibility(Trk2), not_visible, curr_time).
```

```
resume(trk_13,det_1) assign(trk_15,det_0) assign(trk_12,det_2) assign(trk_7,det_3)
assign(trk_8,det_4) assign(trk_3,det_5)
```

```
occurs_at(unhides_from_behind(trk_13,trk_12),268))
```

Similarly, when looking at a slightly more complex scene, like the one depicted in Fig. 10, we get an event sequence describing the interactions happening in the scene:

```
...
 occurs_at(hides_behind(trk_34,trk_16),283)
 occurs_at(unhides_from_behind(trk_34,trk_16),293)
 occurs_at(hides_behind(trk_37,trk_34),296)
 occurs_at(unhides_from_behind(trk_37,trk_34),311)
 ...
```

FIGURE 10: Abducing Event Sequences (Example from MOT Dataset)

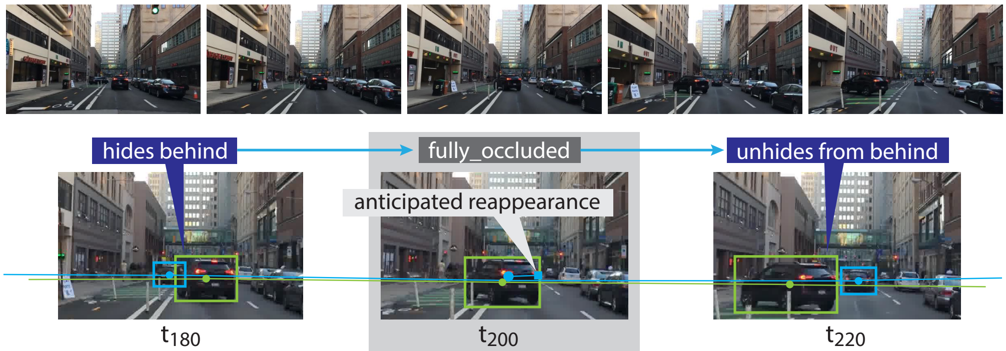


FIGURE 11: Abducing Occlusion to Anticipate Reappearance

This event sequence explains the visuospatial dynamics of the scene and can be used for reasoning about the scene.

4.1.2. *Reasoning about Hidden Entities*

Consider the situation of Fig. 11: a car gets occluded by another car turning left and reappears *in front of* the autonomous vehicle. Using online abduction for abducing high-level interactions of scene objects we can hypothesize that the car got *occluded* and anticipate its reappearance based on the perceived scene dynamics.

The predictions for each track is given by the predicted bounding box, the state in which the track currently is, and the type of the tracked object:

```
trk(trk_3, car). trk_state(trk_3, active).
...
trk(trk_41, car). trk_state(trk_41, active).
...

box2d(trk_3, 660, 460, 134, 102).
...
box2d(trk_41, 631, 471, 40, 47).
...
```

Based on this problem specification for time point 179, the event *hides_behind*(*trk*_41, *trk*_3) is abduced, as there is no detection that could be associated with *trk*_41 and *trk*_3 is partially overlapping with *trk*_41:

```
... occurs_at(hides_behind(trk_41,trk_3),179) ...
```

The abduced explanation together with the object dynamics may then be used for visual reasoning and anticipation of events, which can serve for decision support. Towards this we define a rule stating that a *hidden* object may *unhide* from behind the object it is hidden by and anticipate the time point *t* based on the object *movement* as follows:

```
anticipate(unhides_from_behind(Trk1, Trk2), T) :-
    time(T), curr_time < T,
    holds_at(hidden_by(Trk1, Trk2), curr_time),
    topology(proper_part, Trk1, Trk2),
    movement(moves_out_of, Trk1, Trk2, T).
```

We then interpolate the objects position at time point *t* to predict where the object may *reappear*:

```
point2d(interpolated_position(Trk, T), PosX, PosY) :-
    time(T), curr_time < T, T1 = T-curr_time,
    box2d(Trk1, X, Y,_,_), trk_mov(Trk1, MovX, MovY),
    PosX = X+MovX*T1, PosY = Y+MovY*T1.
```

For the occluded car in our example we get the following prediction for time *t* and position *x*, *y*:

```
anticipate(unhides_from_behind(trk_41, trk_2), 202)
point2d(interpolated_position(trk_41, 202), 738, 495)
```

Based on this prediction we can then define a rule that gives a warning if a hidden entity may reappear in front of the vehicle, which could be used by the control mechanism, e.g., to adapt driving and slow down in order to keep safe distance:

```
warning(hidden_entity_in_front(Trk1, T)) :-
    time(T), T-curr_time < anticipation_threshold,
    anticipate(unhides_from_behind(Trk1, _), T),
    position(in_front, interpolated_pos(Trk1, T)).
```

4.2. Empirical Performance Analysis

For online sensemaking, evaluation focusses on accuracy of abduced motion tracks, real-time performance, and the tradeoff between performance and accuracy. Our evaluation uses the **KITTI** *object tracking dataset* [39], which is a community established benchmark dataset for autonomous cars: it consists of 21 training and 29 test scenes, and provides accurate track annotations for 8 object classes (e.g., car, pedestrian, van, cyclist). We also evaluate tracking results using the more general cross-domain **Multi-Object Tracking** (MOT) dataset [54] established as part of the *MOT Challenge*; We evaluate on MOT 2017 consisting of 7 training and 7 test scenes which are highly unconstrained videos filmed with both static and moving cameras, and

BENCHMARK	MOTA↑	MOTP↑	ML↓	MT↑	FP↓	FN↓	ID sw.↓	Frag.↓
KITTI tracking – *Cars* (8008 frames, 636 targets)								
– baseline	45.72 %	76.89 %	19.14 %	23.04 %	785	11182	1097	1440
– with Abd.	**50.5 %**	74.76 %	20.21 %	23.23 %	1311	10439	165	490
KITTI tracking – *Pedestrians* (8008 frames, 167 targets)								
– baseline	28.71 %	71.43 %	26.94 %	9.58 %	1261	6119	539	833
– with Abd.	**32.57 %**	70.68 %	22.15 %	14.37 %	1899	5477	115	444
MOT 2017 (5316 frames, 546 targets)								
– baseline	41.4 %	88.0 %	35.53 %	16.48 %	4877	60164	779	741
– with Abd.	**46.2 %**	87.9 %	31.32 %	20.7 %	5195	54421	800	904
MOT 2020 (8931 frames, 2332 targets)								
– baseline	49.5 %	87.1 %	17.79 %	18.19 %	5271	531529	36560	39874
– with Abd.	**50.7 %**	87.2 %	18.65 %	17.16 %	4120	537427	17658	38346

TABLE 7: Evaluation of Tracking Performance; accuracy (MOTA), precision (MOTP), mostly tracked (MT) and mostly lost (ML) tracks, false positives (FP), false negatives (FN), identity switches (ID Sw.), and fragmentation (Frag.).

MOT 2020 consisting of 4 training and 4 test scenes filmed in crowded environments. We evaluate on the available groundtruth for training scenes of both KITTI using YOLOv3 detections, and MOT17 / MOT20 using the provided faster RCNN (Region Based Convolutional Neural Network [61]) detections.

4.2.1. Evaluating Object Tracking

For evaluating *accuracy* (MOTA) and *precision* (MOTP) of abduced object tracks we follow the Clear MOT [8] evaluation schema.

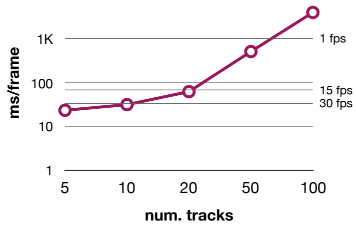
- *MOTA* describes the accuracy of the tracking, taking into account the number of missed objects / false negatives (FN), the number of false positives (FP), and the number of miss-matches (MM).
- *MOTP* describes the precision of the tracking based on the distance of the hypothesised track to the ground truth of the object it is associated to.

These metrics are used to assess how well the generated visual explanations describe the low-level motion in the scene.

Results (Table 7) show that jointly abducing high-level object interactions together with low-level scene dynamics increases the accuracy of the object tracks, i.e, we consistently observe an improvement of about 5% on KITTI and MOT 2017. On KITTI MOTA improves from 45.72% to 50.5% for *cars* and 28.71% to 32.57% for *pedestrians*, and on MOT 2017 it improves from 41.4% to 46.2%. On MOT 2020 we still observe an improvement of 1.2% from 49.5% to 50.7%. This relatively small improvement is mainly because of the different nature of the dataset, i.e., the focus on crowded scenes filmed from a slightly above perspective, which leads to only few targets that get fully occluded by others, and thus there are fewer corrected tracks when the using abductive sensemaking compared to the scenes in KITTI and MOT 2017.

DETECTOR	Recall	MOTA	MOTP	fps_{det}	fps_{abd}
YOLOv3	0.690	50.5 %	74.76 %	45	33.9
SSD	0.599	30.63 %	77.4 %	8	46.7
FRCNN	0.624	37.96 %	72.9 %	5	32.0

FIGURE 12: Online Performance; performance for pretrained detectors (DET.) on the *'cars'* class of KITTI dataset



No. of TRACKS	$ms/frame$	fps
5	23.33	42.86
10	31.36	31.89
20	62.08	16.11
50	511.83	1.95
100	3996.38	0.25

FIGURE 13: Scalability; processing time relative to the no. of tracks on synthetic dataset.

4.2.2. Online Performance and Scalability

Performance of online abduction is evaluated with respect to its real-time capabilities.[7] **(1).** We compare the time & accuracy of online abduction for state of the art (real-time) detection methods: YOLOv3, SSD [50], and Faster RCNN [61] (Fig. 12). **(2).** We evaluate scalability of the ASP based abduction on a synthetic dataset with controlled number of tracks and percentage of overlapping tracks per frame. Results (Fig. 13) show that online abduction can perform with above 30 frames per second for scenes with up to 10 highly overlapping object tracks, and more than 50 tracks with 1fps (for the sake of testing, it is worth noting that even for 100 objects per frame it only takes about an average of 4 secs per frame). Importantly, for realistic scenes such as in the KITTI dataset, abduction runs realtime at 33.9fps using YOLOv3, and 46.7 using SSD with lower accuracy but providing good precision.

4.2.3. Discussion of Empirical Results

Results show that integrating high-level abduction and object tracking improves the resulting object tracks and reduce the noise in the visual observations. For the case of online visual sense-making, ASP based abduction provides the required performance: even though the complexity of ASP based abduction increases quickly, with large numbers of tracked objects the framework can track up to 20 objects simultaneously with $30fps$ and achieve real-time performance on the KITTI benchmark dataset. It is also important to note that the tracking approach in this paper is based on *tracking by detection* using a naive measure, i.e, the IoU (Sec. 3.1; Step 1), to associate observations and tracks, and it is not using any visual information in the prediction or association step. Naturally, this results in a lower accuracy, in particular when used with noisy detections and when tracking fast moving objects in a benchmark dataset such as KITTI. That said, due to the modularity of the implemented framework, extensions with different methods for predicting

[7]Evaluation using a dedicated Intel Core i7-6850K 3.6GHz 6-Core Processor, 64GB RAM, and a NVIDIA Titan V GPU 12GB.

motion (e.g., using particle filters or optical flow based prediction) are straightforward: i.e., improving tracking is not the aim of our research.

5. DISCUSSION AND RELATED WORK

Answer Set Programming is now widely used as a foundational declarative language and robust methodology for a range of (non-monotonic) knowledge representation and reasoning tasks [23, 62, 37, 36, 38]. With ASP as a foundation, and driven by semantics, commonsense and explainability [30, 29], this research aims to bridge the gap between high-level formalisms for logical visual reasoning (e.g., by abduction) and low-level visual processing by tightly integrating semantic abstractions of *space and change* with their underlying numerical representations. More broadly, this goal is pursued within the larger agenda of *cognitive vision and perception* [12], which is an emerging line of research bringing together a novel & unique combination of methodologies from Artificial Intelligence, Vision and Machine Learning, Cognitive Science and Psychology, Visual Perception, and Spatial Cognition and Computation. Research in cognitive vision and perception addresses visual, visuospatial and visuo-locomotive perception and interaction from the viewpoints of language, logic, spatial cognition and artificial intelligence [74, 67, 72, 70, 71]. In this broader context, the principal motivation and developmental goal of this research follows a one-point agenda, namely:

> to develop a systematic, general, and modular integration of (methods in) Computer Vision and AI, particularly emphasising the integration of high-level knowledge representation and reasoning techniques with low-level (i.e., quantitatively) based visual computing techniques (which in the present scientific status quo are primarily driven by end-to-end, black-box deep learning pipelines).

The integration of Vision and AI addressed in our research is motivated by the need to realise human-centred criteria pertinent to the design and implementation of high-level visual sensemaking technology, e.g., within autonomous driving systems where such criteria emanating from standardisation and regulation considerations are of utmost priority. Although this paper selectively focusses on the needs and challenges of active / online sensemaking in autonomous driving, the generality and modularly of the developed framework ensures foundational applicability in diverse applied contexts requiring perception, interaction and control; e.g., a case in point here being the fact that the demonstrated application and evaluation also directly function with general datasets such as MOT concerned with moving objects (Sec 4). Of at least equal importance are the modularity and elaboration tolerance of the framework, enabling seamless integration and experimention with advances in fast evolving computer vision methods, as well as experimenting with different forms of formal methods for *reasoning about space, actions, and change* [10, 11] that could either be embedded directly within answer set programming, or possibly be utilised independently as part of other declarative frameworks for knowledge representation and reasoning.

Perception and Abduction: A KR Perspective. Within KR, the significance of high-level (abductive) explanations in a range of contexts is long established: planning & process recognition [44, 43], vision & abduction [66], probabilistic abduction [19], reasoning about spatio-temporal dynamics [11], reasoning about continuous *spacetime* change [56, 41] etc. Dubba et al. [32]

formalises abductive reasoning in an inductive-abductive loop within inductive logic programming (ILP). Aditya et al. [1] formalise general rules for image interpretation with ASP. Closely related to this research is [73], which uses a two-step approach (with one huge *problem specification*), first tracking and then explaining (and fixing) tracking errors; such an approach is not runtime / realtime capable. Within computer vision research there has recently been an interest to synergise with cognitively motivated methods; in particular, e.g., for perceptual grounding and inference [85], and combining video analysis with textual information for understanding events and answering queries about video data [78].

Perception in Autonomous Driving. The present industrial relevance and market potential[8] of autonomous driving technology can be primarily attributed to recent advances in deep learning driven computer vision. A typical engineering stack for autonomous driving consists of perception, prediction, planning and control modules [87]: perception gives the location, pose of the objects in the world while prediction forecasts the motion of the objects; planning involves creating a trajectory for the motion of the vehicle which is then executed by the controller. In object detection, Tan and Le [77] introduced EfficientDet which achieves order-of-magnitude better efficiency than previous works [60, 61, 50, 77] without any drop in performance. Large datasets and self-supervised methods [25, 88] enable end to end joint learning of flow, depth and camera pose estimation more accurately, exploiting the inherent relation between each other. More specialised research on object detection has investigated specific cases relevant in driving, such as detection of smaller objects [84], partially occluded pedestrians [58] and 3D object detection [83, 49, 89]. Recent advancements in object tracking involves neural methods like [7] which employ a tracking by detection paradigm and predicts the next object position using a simple neural network. Multi-object tracking is also extended to multi-object tracking and segmentation by [79]. Semantic and Instance segementation of the object [26] [90] [86] [76] provides accurate boundaries. Advancements in robust lane detection [42] make it possible to extend automatic lane keeping and lane switching. Recent neural methods estimate visual odometry, ego motion, depth and flow through a set of multi-task learning methodologies. [88] [52] show that depth and ego-motion can be learnt in a joint manner. Flow and depth are also learnt using a multi-task approach [25] [91] [82].

Hybrid Methods to Meet Multi-Faceted Challenges. Critical challenges in driving, e.g., pertaining to perception, prediction, planning and control modules [87], are researched and developed individually which leads to a sub-optimal overall performance. End-to-end driving methodologies [21, 87] are constructed in such a way that the sensor outputs such as images, LiDAR (Light Detection and Ranging) are directly used to predict control signals like steering and acceleration. Furthermore, these methods are generally black-box and are unable to model the complex multi-faceted nature of autonomous driving encompassing dimensions of human fac-

[8] **Industrial initiatives in autonomous driving**. Autonomous driving research within industry is now well established: there exist cab-sharing companies like Uber and Lyft attempting to replace human drivers with "fully autonomous self-driving" vehicles. Companies such as Baidu, Comma AI and organisations like Udacity are creating an open source platform for various technologies of the self driving stack. Manufacturing giants such as GM, Toyota, Ford, Daimler, Bosch are also taking steps to offer varying levels of autonomy to consumer and industrial vehicles either directly or indirectly; GM acquired Cruise Automation while Toyota has invested in and collaborates with pony.ai. Ford and Volkswagen has partnered with Argo AI to bring self driving capabilities to their respective vehicles. Waymo, a subsidiary of Alphabet has already deployed autonomous ride-sharing operations in two cities. Last, but not the least, is Tesla with its competitive advantage of already having close to 500000 cars on the road collecting data with (Level 2 assistance) AutoPilot enabled.

tors and usability, (natural) roadside multimodal interaction [46, 45] etc, or support the range of human-centred AI considerations related to declarative explainability, queryability etc that have been the principal impulse underlying the aims of the methods developed in this paper.

Our research achieves a systematic integration of KR and Vision methods hitherto developed, evaluated, and applied in completion isolation of one another; we believe that our resulting framework can serve as a one possible interpretation and exemplar for the neurosymbolic integration of relational AI and neural (visual) feature detectors. Furthermore, it offers a novel potential for a multifaceted but integrated applied evaluation and benchmarking of visual sensemaking technologies: e.g., it is common practice within computer vision research to evaluate and benchmark visual computing capabilities, e.g., for object detection, tracking, using absolute performance benchmarks either solely or primarily centred on incremental improvements in accuracy. Naturally, this is necessary for fundamental progress in vision research, but such an evaluation metric misses out on other crucial requirements as they pertain to human-centred AI considerations in applications domains such as autonomous driving. For instance, in light of ethically driven standardisation and regulatory considerations (Section 1), this research has been motivated and directly addresses interpretability and explainability challenges (**C1 – C4**):

C1. *Active visual sensemaking*, e.g., involving (real-time) commonsense visuospatial abduction and (simulated) prediction of grounded percepts

C2. *Posthoc analysis of quantitative archives*, e.g., requiring semantic search / retrieval / visualisation for diagnosis, dispute settlement, inspection

C3. *Natural human-machine interaction*, e.g., involving natural language interfaces for (explanatory) communication between vehicle and passengers (or other stakeholders)

C4. *Standardisation* for vehicular licensing & validation, e.g., involving creation of diverse, naturalistic datasets usable in testing of autonomous vehicle performance; how to access the quality and distribution of training datasets utilised? (Sec 6)

Our research, by its integrative approach, makes it possible to explicitly address "human-centred interpretability and explainability challenges" such as in (C1–C4) for autonomous driving systems at the practical level of methods and tools. This is especially beneficial and timely since not everything in autonomous vehicles is about realtime control / decision-making; several human-machine interaction requirements (e.g., for interpretable diagnostic communication, universal design) also exist. The Federal Ministry of Transport and Digital Infrastructure in Germany (BMVI) has taken a lead in eliciting 20 key propositions (with possible legal implications) for the fulfilment of ethical commitments for automated and connected driving systems [20]. The BMVI report highlights a range of factors pertaining to safety, utilitarian considerations, human rights, statutory liability, technological transparency, data management and privacy etc. We claim that what appears as spectrum of complex challenges (in autonomous driving) that may possibly delay technology adoption is actually rooted to one fundamental methodological consideration that needs to be prioritised, namely: the design and implementation of human-centred technology based on a *confluence* of techniques and perspectives from AI+ML, Cognitive Science & Psychology, Human-Machine Interaction, and Design Science. Like in many applications of AI, such an integrative approach has so far not been explored also within autonomous driving research.

6. SUMMARY AND OUTLOOK

We have developed a novel neurosymbolic abduction-driven *online* (i.e., realtime, incremental) visual sensemaking framework: general, systematically formalised, modular, and fully implemented. Integrating robust state-of-the-art methods in *knowledge representation* and *computer vision*, the framework has been evaluated and demonstrated with established community benchmarks. We highlight application prospects of semantic vision for autonomous driving, a domain of emerging and long-term significance for research in Artificial Intelligence and Machine Learning. From the applied viewpoint of autonomous driving, our work is motivated by interpretability and explainability benchmarks (e.g., in active visual sensemaking, posthoc analysis, natural human-machine interaction, standardisation for licensing & validation; Sec 4.1) that go far beyond basic considerations in contemporary autonomous driving research, namely: *how fast to drive and which way to steer*, and testing performance by *clocking mileage* alone by the use of deep learning based methods in training and testing phases.

Technical Extensions

Our development of a systematic, modular, and general visual sensemaking methodology opens up several possibilities for further technical developments / extensions:

- *Commonsense*. Specialised commonsense theories about multi-sensory integration, multi-agent belief merging, incorporation of contextual knowledge and situational norms within the declarative framework of ASP merits individual strands of further development.
- *Tracking by detection*. Given the modularity of the developed framework, incorporating and experimenting with specialised / emerging low-level visual computing methods becomes feasible with relative ease. For instance, in this paper we have not attempted to develop a new tracking algorithm as such; instead, one of our aims has been to showcase the manner in which perceptual sensemaking by visual abduction can be integrated into a standard "tracking by detection" paradigm, which is most widely used approach in state of the art tracking (Sec 5). Nevertheless, extensions and variations of this approach deserve further investigation where tracking itself takes a centre-stage.
- *Uncertainty*. The present work handles the uncertainty involved in low-level object tracking using a naive approach, which suffices for the present purposes, i.e., a full-scale systematic formalisation of a probabilistic model has not been attempted herein. However, handling uncertainty calls for its systematic treatment, e.g., requiring either integrating a declarative probabilistic model directly within the answer set programming framework, or possibly independently as a separately module. One seemingly natural approach towards this would be to explore possibilities with probabilistic ASP [48].

Towards a Dataset: Reasoning and Scenario Visuospatial Complexity Coverage. The application demonstrations of this paper have been conducted in the backdrop of select safety-critical situations (Table 6; e.g., Fig. 7), without aiming to achieve an exhaustive collection (if at all it is even possible to be comprehensive in this respect). The scenarios and corresponding safety-criticality are exemplary, with the selections emanating from a behavioural study of human-factors in everyday driving situations, and safety criticality determined based on analysis of empirical data about roadside accidents / hazardous situations from publicly available data published in accident research reports, e.g., by the German Insurance Association ("*Unfallforschung der Versicherer*") [35]. Work is presently in progress to develop novel benchmark datasets (in synergy with behavioural human studies; refer below) that centralise range and distribution vis-vis commonsense explainability and visuospatial complexity ([45, 46]) criteria classes within a dataset, as opposed to merely collecting accumulating "mileage" / "big data".

Human-Factors in Autonomous Driving:
A Cognitive Methodology Combining Behavioural and Computational Approaches

In addition to continuing (aforediscussed) technical developments in computational cognitive vision pertaining to the integration of "vision & AI", our ongoing focus is to develop a novel dataset emphasising (visuospatial) semantics and (commonsense) explainability. For instance, we develop a methodology —focussing on *visuospatial complexity* [46] of stimuli and *multimodal interactions* [45] in ecologically valid naturalistic [3, 59] driving conditions— for establishing human-centred benchmarks and corresponding testing & validation datasets for visual sensemaking primarily from a human cognitive factors viewpoint. Our particular focus here is on embodied multimodal interactions (e.g., gestures, joint attention, visual search complexity) amongst drivers, pedestrians, cyclists etc under ecologically valid naturalistic conditions. This initiative is driven by bottom-up interdisciplinary research –encompassing AI, Psychology, HCI, and Design– for the study of driving behaviour particularly in diverse low-speed, complex urban environments possibly with unstructured traffic. Such interdisciplinary studies *–at the confluence of Cognition, AI, Interaction, and Design–* are needed to better appreciate the complexity and spectrum of varied human-centred challenges in autonomous driving, and demonstrate the significance of integrated vision & AI solutions [12] in those contexts.

ACKNOWLEDGEMENTS.

We thank the reviewers at IJCAI 2019 [74] for their constructive feedback and support of our work; all reviewer suggestions that could not be included in the conference length paper have been fully incorporated in the present article. We are also grateful to the anonymous reviewers and editors at the AIJ journal whose comments have helped us further improve the (final published) paper; we remain especially appreciative for their timely service during the special times of 2020.

We acknowledge partial funding by the German Research Foundation (DFG) via the Collaborative Research Center 1320 EASE – "Everyday Activity Science and Engineering" (www.ease-crc.org) project:

Spatial Reasoning in Everyday Activity., Number 329551904.

http://gepris.dfg.de/gepris/projekt/374123335

APPENDICES.

AppendixA. SELECT ANSWER SET PROGRAMMING CODE

We envisage that all applicable electronic material (data sets, programs, videos....) will be published as a supplement in an archival format in due course. As for this appendix, select code snippets in support of the examples in the paper are included; a full implementation will be linked and released (also including a light-weight execution environment to be determined) upon final publication of the paper independent of the proposed supplementary publication.

AppendixA.1. Abduction Based Association

Following the generate and test paradigm of ASP, **choice rules** are used to generate all assignments between detections and tracks to resulting on all possible assignments; assignments are tested using **integrity constraints**.

- Choice rules for generating assignment actions generate the set of assignments actions for all tracks and all detections; for example:

```
1{
    assign(Trk, Det): det(Det, _, _);
    end(Trk);
    ignore_trk(Trk);
    halt(Trk);
    resume(Trk, Det): det(Det, _, _)
}1
:- trk(Trk, Trk_Type).

1{
    assign(Trk, Det): trk(Trk, _);
    start(Det);
    ignore_det(Det);
    resume(Trk, Det): trk(Trk, _)
}1
:- det(Det, Det_Type, Conf).
```

- Generated assignments are tested based on (spatio-temporal) constraints for each assignment action. Assignments not consistent with these constraints are eliminated from the set of answers using integrity constraints:

```
:- assign(Trk, Det), not assignment_constraints(Trk, Det).
:- start(Det), not start_constraints(Det).
:- end(Trk), not trk_state(Trk, halted).
:- ignore_trk(Trk), not trk_state(Trk, halted).
:- halt(Trk), not trk_state(Trk, active).
:- resume(Trk, Det), not resume_constraints(Trk, Det).

assignment_constraints(Trk, Det) :-
    trk(Trk, Trk_Type), det(Det, Det_Type, Conf),
    trk_state(Trk, active), match_type(Trk_Type, Det_Type),
    Conf > conf_thresh_assign,
    iou(Trk, Det, IOU), IOU > iou_thresh.
```

```
resume_constraints(Trk, Det) :-
    trk(Trk, Trk_Type),
    det(Det, Det_Type, Conf), Conf > conf_thresh_resume,
    match_type(Trk_Type, Det_Type),
    trk_state(Trk, halted).

start_constraints(Det) :-
    det(Det, _, Conf), Conf > conf_thresh_new_track,
    size(bigger, Det, size_threshold).
```

This results in the set of all possible assignments, which further gets optimized based on **optimization statements** in AppendixA.4.

AppendixA.2. Abducible High-Level Events

Event hypotheses with respect to background fluents and events are generated to explain assignment actions.

- Functional fluent *visibility* of a track can be *fully_visible*, *partially_visible*, or *not_visible*.

```
fluent(visibility(Trk)) :- trk(Trk, _).

possVal(visibility(Trk), fully_visible) :- trk(Trk, _).
possVal(visibility(Trk), partially_visible) :- trk(Trk,_).
possVal(visibility(Trk), not_visible) :- trk(Trk, _).
```

- Boolean fluent *hidden_by* for two tracks can be *true* or *false*.

```
fluent(hidden_by(Trk1, Trk2)) :- trk(Trk1,_), trk(Trk2,_).

possVal(hidden_by(Trk1, Trk2), true) :- trk(Trk1, _), trk(Trk2, _).
possVal(hidden_by(Trk1, Trk2), false) :- trk(Trk1, _), trk(Trk2, _).
```

- Boolean fluent *clipped* for a track can be *true* or *false*.

```
fluent(clipped(Trk)) :- trk(Trk,_).

possVal(clipped(Trk), true) :- trk(Trk, _).
possVal(clipped(Trk), false) :- trk(Trk, _).
```

- Fluents corresponding to all tracks and pairs of tracks are initialised as follows: all tracks are initialised as fully visible, not hidden by another track, and not clipped (however, note that it can be the case that events occurring with the start of a track have an effect on initialised fluent values, e.g., an event for a track starting partially occluded).

```
holds_at(clipped(Trk), false, mintime) :- trk(Trk, _).
holds_at(hidden_by(Trk1, Trk2), false, mintime) :- trk(Trk1, _), trk(Trk2, _).
holds_at(visibility(Trk), fully_visible, mintime) :- trk(Trk, _).
```

- Events and causal effects are defined to describe changes in the fluents as effects of events occurring in the world. Here we show examples for the events *hides_behind* and *missing_detections*.

The event *hides_behind* is defined on two tracks as follows:

```
event(hides_behind(Trk1,Trk2)) :- trk(Trk1,_),trk(Trk2,_).
```

One object hiding behind another object causes the visibility fluent for the hidden object to change its value to *not_visible*. Further the fluent *hidden_by* for the two tracks changes its value to *true*.

```
causesValue(hides_behind(Trk1, Trk2), visibility(Trk1), not_visible, T) :-
trk(Trk1,_), trk(Trk2,_), time(T).

causesValue(hides_behind(Trk1, Trk2), hidden_by(Trk1, Trk2), true, T) :-
trk(Trk1,_), trk(Trk2,_), time(T).
```

The event *missing_detections* is defined on a single track as follows.

```
event(missing_detections(Trk)) :- trk(Trk,_).

causesValue(missing_detections(Trk), clipped(Trk), true, T) :-
trk(Trk,_), time(T).
```

AppendixA.3. Abducing High-Level Events Explaining Assignments

Possible explanations are generated using choice rules for explaining association actions, i.e., for each association a possible explanation in terms of high-level events is generated based on preconditions and causal effects. Here we show examples for the events *hides_behind* and *missing_detections*.

- Choice rule (snippet) for explaining halted tracks:

A track can be halted because it is hiding behind another track, or there are missing detections within the track.

```
1{
    occurs_at(hides_behind(Trk, Trk2), curr_time): trk(Trk2,_);
    occurs_at(missing_detections(Trk), curr_time)
}1
:- halt(Trk).
```

- Constraints for events are defined using integrity constraints for each event:

Integrity constraint for event *hides_behind* can not occur if *poss(hides_behind(_, _))* is not true.

```
:- occurs_at(hides_behind(Trk1, Trk2), curr_time), not poss(hides_behind(Trk1, Trk2)).
```

- The event *hides_behind* is possible if the tracks are overlapping and both tracks visible.

```
poss(hides_behind(Trk1, Trk2)) :-
    trk(Trk1, _), trk(Trk2, _),
    position(overlapping_top, Trk1, Trk2),
    not holds_at(visibility(Trk1), not_visible, curr_time),
    not holds_at(visibility(Trk2), not_visible, curr_time).
```

- Integrity constraint for event *missing_detections*.

```
:- occurs_at(missing_detections(Trk), curr_time), not poss(missing_detections(Trk)).
```

- The event *missing_detections* is possible if the track is not clipped and it is visible.

```
poss(missing_detections(Trk)) :-
    holds_at(clipped(Trk), false, curr_time),
    not holds_at(visibility(Trk), not_visible, curr_time).
```

AppendixA.4. Optimisation

- Finding best fitting hypothesis on assignments and high-level events is achieved using ASP optimisation statements as follows:

— Matching likelihood is maximised to ensure matching of best fitting detections to tracks, i.e., here maximising IoU between bounding rectangles of predicted tracks and the detections:

```
matching_likelihood(Trk, Det, IOU) :- det(Det, _, _), trk(Trk, _), iou(Trk, Det, IOU).
```

```
A#maximize {(ML)@10,Trk,Det : assign(Trk, Det), matching_likelihood(Trk, Det, ML)}.
```

— Maximising assignment of detections to tracks to avoid segmented tracks, i.e., assign detections to tracks whenever possible:

```
A#maximize {1@10, Trk, Det: assign(Trk, Det)}.
```

— Resume tracks if possible; start / end tracks if resuming is not possible:

```
A#minimize {1@2, Trk, Det: resume(Trk, Det)}.

A#minimize {5@2,Trk: end(Trk)}.
A#minimize {5@2,Det: start(Det)}.
```

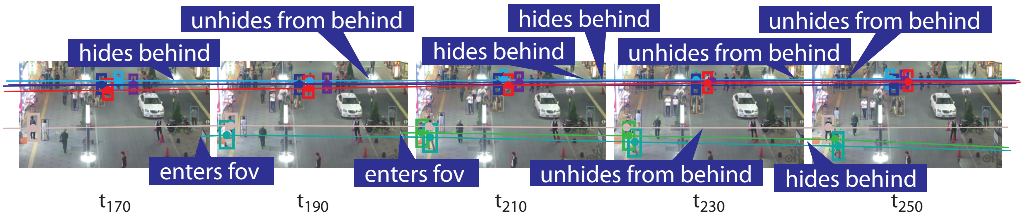
— Only if no other explanation can be found, tracks and detections are ignored:

```
A#minimize {10@3,Det: ignore_det(Det)}.
A#minimize {10@3,Trk: ignore_trk(Trk)}.
```

AppendixB. ADDITIONAL EXAMPLES

AppendixB.1. Occlusion Example

Abduced event sequence for scene 04 from the MOT 2017 benchmark, involving people moving in a crowded environment, with various occlusions.




```
...
occurs_at(enters_fov(trk_30),172)
occurs_at(hides_behind(trk_14,trk_22),188)
occurs_at(hides_behind(trk_16,trk_30),191)
occurs_at(enters_fov(trk_32),203)
occurs_at(unhides_from_behind(trk_14,trk_22),205)
occurs_at(hides_behind(trk_14,trk_20),222)
occurs_at(hides_behind(trk_8,trk_22),230)
occurs_at(unhides_from_behind(trk_16,trk_32),238)
occurs_at(hides_behind(trk_32,trk_30),239)
occurs_at(unhides_from_behind(trk_14, trk_20),245)
occurs_at(unhides_from_behind(trk_8,trk_22),250)
...
```

FIGURE B.14: Abduced events for scene MOT17-04 between time point 270 and time point 310.

AppendixB.2. **Results for Select (Complete) Scenes**

The following are results for select scenes from the datasets being used in the evaluation (Sec 4): KITTIMOD, MOT, and safety-criticality set of scenarios developed as part of this work. For lack of space, we only choose to illustrate one select frames per sec of input stimuli:

- Figure B.15: Scene 20 from KITTIMOD [39] tracking dataset
- Figure B.16: Scene 02 from the MOT Challenge [54]
- Figure B.17: Scene from safety-critical scenario dataset (Sec 4.1.2)

FIGURE B.15: Complete example scene form KITTIMOD tracking benchmark. High traffic highway situation including high and low speed driving.

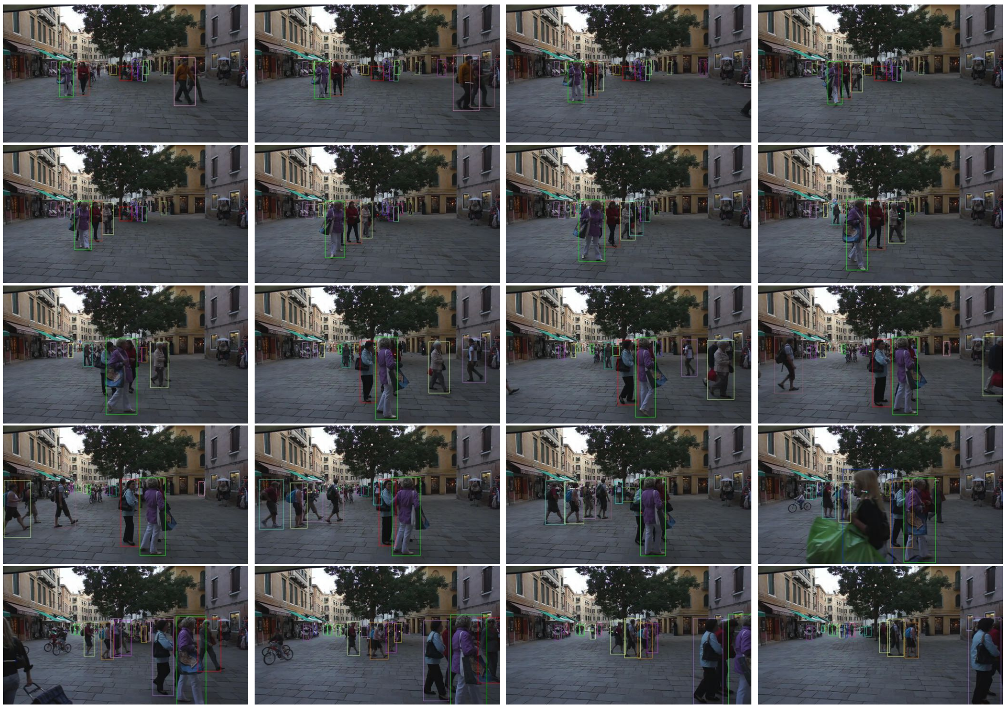


FIGURE B.16: Complete example scene from MOT16 benchmark dataset for people tracking.

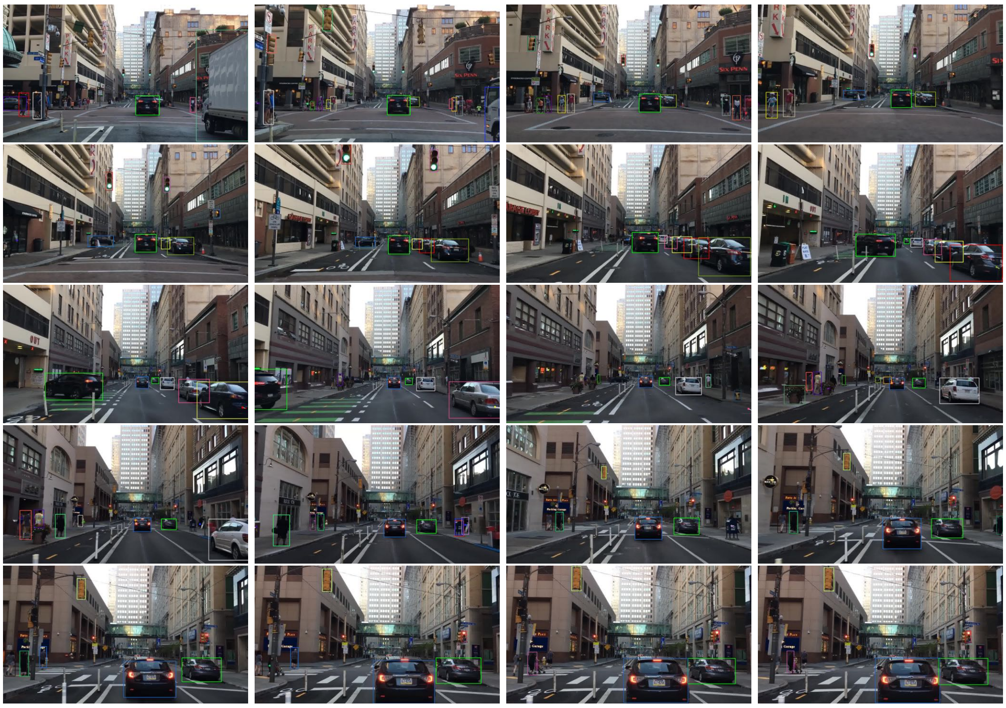


FIGURE B.17: Tracking results for the complete scene of the occlusion example (Fig 11; Section 4.1.2) involving tracking of cars, pedestrians, and traffic lights

AppendixC. EXAMPLE DATA

The problem specification for each time point t, which is the input data for the answer set programming based abduction, is generated online based on the visual stimuli; because of the size of the data (visual observations, predictions, and matching likelihood for each frame of the video) we only include a snippet for one frame to illustrate the nature of the data.

Example problem specification ($< \mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t >$) generated for KITTI 0020, time point 79.

```
A#const curr_time=79.
```

$\mathcal{VO}_{79}$ — Spatial entities of detected objects as bounding boxes:

```
%% detections
det(det_0, car, 99).
det(det_1, car, 99).
det(det_2, car, 99).
det(det_3, car, 99).
det(det_4, car, 95).
det(det_5, car, 91).
det(det_6, car, 84).
det(det_7, car, 75).
det(det_8, car, 72).
det(det_9, car, 46).
det(det_10, car, 40).
det(det_11, car, 25).
det(det_12, car, 22).
det(det_13, truck, 52).
det(det_14, truck, 52).

% boxes for detections
box2d(det_0, 0, 189, 208, 119).
box2d(det_1, 697, 187, 105, 68).
box2d(det_2, 220, 178, 215, 138).
box2d(det_3, 401, 183, 89, 72).
box2d(det_4, 640, 179, 38, 28).
box2d(det_5, 520, 179, 27, 23).
box2d(det_6, 473, 182, 39, 33).
box2d(det_7, 588, 179, 30, 22).
box2d(det_8, 494, 184, 29, 29).
box2d(det_9, 557, 176, 11, 14).
box2d(det_10, 475, 173, 28, 18).
box2d(det_11, 422, 174, 39, 13).
box2d(det_12, 453, 176, 24, 12).
box2d(det_13, 586, 174, 32, 22).
box2d(det_14, 579, 172, 21, 20).
```

$\mathcal{P}_{79}$ — Spatial entities of predicted tracks for time-point 79 as bounding boxes:

```
% tracks and track states
trk(trk_0, car).
trk_state(trk_0, halted).
trk(trk_1, car).
trk_state(trk_1, active).
```

```
trk(trk_4, car).
trk_state(trk_4, active).
trk(trk_5, car).
trk_state(trk_5, active).
trk(trk_6, car).
trk_state(trk_6, active).
trk(trk_7, car).
trk_state(trk_7, active).
trk(trk_9, car).
trk_state(trk_9, halted).
trk(trk_11, car).
trk_state(trk_11, active).
trk(trk_12, car).
trk_state(trk_12, active).
trk(trk_13, car).
trk_state(trk_13, active).

% boxes for tracks
box2d(trk_0, -42, 227, 249, 159).
box2d(trk_1, 698, 186, 102, 68).
box2d(trk_4, 590, 179, 26, 21).
box2d(trk_5, 639, 179, 39, 27).
box2d(trk_6, 245, 187, 182, 115).
box2d(trk_7, 495, 181, 27, 31).
box2d(trk_9, 319, 184, 54, 41).
box2d(trk_11, -26, 188, 235, 113).
box2d(trk_12, 404, 181, 85, 70).
box2d(trk_13, 522, 179, 23, 22).
```

$\mathcal{ML}_{79}$ — Matching likelihood for pairs of tracks and detections at time point 79 given by the IoU between them.

```
% IoU for overlapping tracks and detections
iou(trk_0,det_0,34625).
iou(trk_11,det_0,83400).
iou(trk_1,det_1,96879).
iou(trk_6,det_2,71071).
iou(trk_9,det_2,7556).
iou(trk_12,det_2,6388).
iou(trk_6,det_3,7330).
iou(trk_12,det_3,90971).
iou(trk_5,det_4,94824).
iou(trk_7,det_5,4551).
iou(trk_13,det_5,84391).
iou(trk_7,det_6,30697).
iou(trk_12,det_6,8403).
iou(trk_4,det_7,84365).
iou(trk_7,det_8,86273).
iou(trk_13,det_8,1145).
iou(trk_7,det_10,5146).
iou(trk_12,det_10,2138).
iou(trk_12,det_11,3087).
iou(trk_12,det_12,2341).
iou(trk_4,det_13,51257).
iou(trk_4,det_14,13495).
```

Abduced Event Sequence for time point 79 (snippet for 10 time points)

```
...
occurs_at(missing_detections(trk_9),69)
occurs_at(missing_detections(trk_0),70)
occurs_at(missing_detections(trk_7),72)
occurs_at(noise(trk_10),73)
occurs_at(recover(trk_7),73)
```

Bibliography

[1] S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos. Visual commonsense for scene understanding using perception, semantic parsing and reasoning. In *2015 AAAI Spring Symposium Series*, 2015.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983. ISSN 0001-0782.

[3] M. Angrosino. *Naturalistic Observation*. Qualitative essentials. Left Coast Press, 2007. ISBN 9781598740592.

[4] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018. doi: 10.1038/s41586-018-0637-6. URL https://doi.org/10.1038/s41586-018-0637-6.

[5] P. Balbiani, J. Condotta, and L. F. del Cerro. A new tractable subclass of the rectangle algebra. In T. Dean, editor, *IJCAI 1999, Sweden*, pages 442–447. Morgan Kaufmann, 1999.

[6] B. Bennett, A. G. Cohn, P. Torrini, and S. M. Hazarika. A foundation for region-based qualitative geometry. In *Proceedings of the 14th European Conference on Artificial Intelligence*, pages 204–208, 2000.

[7] P. Bergmann, T. Meinhardt, and L. Leal-Taixé. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[8] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 2008. ISSN 1687-5281. doi: 10.1155/2008/246309. URL https://doi.org/10.1155/2008/246309.

[9] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. doi: 10.1109/ICIP.2016.7533003.

[10] M. Bhatt. Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning. In *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*. IGI Global, USA, 2012. ISBN ISBN13: 9781616928681.

[11] M. Bhatt and S. W. Loke. Modelling dynamic spatial systems in the situation calculus. *Spatial Cognition & Computation*, 8(1-2):86–130, 2008. doi: 10.1080/13875860801926884. URL https://doi.org/10.1080/13875860801926884.

[12] M. Bhatt. and J. Suchan. Cognitive vision and perception: Deep semantics integrating AI and vision for (declarative) reasoning about space, action, and motion. In *24th European Conference on Artificial Intelligence (ECAI)*, Santiago de Compostela, Spain, 2020.

[13] M. Bhatt and J. O. Wallgrün. Geospatial narratives and their spatio-temporal dynamics: Commonsense reasoning for high-level analyses in geographic information systems. *ISPRS Int. J. Geo Inf.*, 3(1):166–205, 2014. doi: 10.3390/ijgi3010166. URL https://doi.org/10.3390/ijgi3010166.

[14] M. Bhatt, H. W. Guesgen, S. Wölfl, and S. M. Hazarika. Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. *Spatial Cognition & Computation*, 11(1):1–14, 2011. doi: 10.1080/13875868.2010.548568. URL https://doi.org/10.1080/13875868.2010.548568.

[15] M. Bhatt, J. H. Lee, and C. P. L. Schultz. CLP(QS): A declarative spatial reasoning framework. In M. J. Egenhofer, N. A. Giudice, R. Moratz, and M. F. Worboys, editors, *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings*, volume 6899 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2011. doi: 10.1007/978-3-642-23196-4_12. URL https://doi.org/10.1007/978-3-642-23196-4_12.

[16] M. Bhatt, C. Schultz, and C. Freksa. The 'Space' in Spatial Assistance Systems: Conception, Formalisation and Computation. In T. Tenbrink, J. Wiener, and C. Claramunt, editors, *Representing space in cognition: Interrelations of behavior, language, and formal models. Series: Explorations in Language and Space*. 978-0-19-967991-1, Oxford University Press, 2013.

[17] M. Bhatt, J. Suchan, and C. P. L. Schultz. Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation. In M. A. Finlayson, B. Fisseni, B. Löwe, and J. C. Meister, editors, *2013 Workshop on Computational Models of Narrative, CMN 2013, August 4-6, 2013, Hamburg, Germany*, volume 32 of *OASICS*, pages 24–29. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013. doi: 10.4230/OASIcs.CMN.2013.24. URL https://doi.org/10.4230/OASIcs.CMN.2013.24.

[18] M. Bhatt, J. Suchan, and S. Vardarajan. Deep semantics for explainable visuospatial intelligence : Perspectives on integrating commonsense spatial abstractions and low-level neural features. In *Proceedings of the 2019 International Workshop on Neural-Symbolic Learning and Reasoning : Annual workshop of the Neural-Symbolic Learning and Reasoning Association*, 2019. URL https://www.researchgate.net/publication/333480472_Deep_Semantics_for_Explainable_Visuospatial_Intelligence_Perspectives_on_Integrating_Commonsense_Spatial_Abstractions_and_Low-Level_Neural_Features.

[19] J. Blythe, J. R. Hobbs, P. Domingos, R. J. Kate, and R. J. Mooney. Implementing weighted abduction in markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, USA, 2011. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2002669.2002676.

[20] BMVI. Report by the ethics commission on automated and connected driving., bmvi: Federal ministry of transport and digital infrastructure, germany, 2018. URL https://www.bmvi.de/goto?id=349482.

[21] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016. URL http://arxiv.org/abs/1604.07316.

[22] J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf2654. URL https://science.sciencemag.org/content/352/6293/1573.

[23] G. Brewka, T. Eiter, and M. Truszczyński. Answer set programming at a glance. *Commun. ACM*, 54(12):92–103, Dec. 2011. ISSN 0001-0782. doi: 10.1145/2043174.2043195.

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

[25] Y. Chen, C. Schmid, and C. Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] E. Davis. Pouring liquids: A study in commonsense physical reasoning. *Artif. Intell.*, 172(12-13):1540–1578, 2008.

[28] E. Davis. How does a box work? a study in the qualitative dynamics of solid objects. *Artif. Intell.*, 175(1):299–345, 2011.

[29] E. Davis. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Intell. Res.*, 59:651–723, 2017. doi: 10.1613/jair.5339. URL https://doi.org/10.1613/jair.5339.

[30] E. Davis and G. Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, 2015. doi: 10.1145/2701413. URL http://doi.acm.org/10.1145/2701413.

[31] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*, Mar. 2020. URL http://arxiv.org/abs/1906.04567. arXiv: 2003.09003.

[32] K. S. R. Dubba, A. G. Cohn, D. C. Hogg, M. Bhatt, and F. Dylla. Learning relational event models from video. *J. Artif. Intell. Res. (JAIR)*, 53:41–90, 2015. doi: 10.1613/jair.4395. URL http://dx.doi.org/10.1613/jair.4395.

[33] M. Eppe and M. Bhatt. Approximate postdictive reasoning with answer set programming. *J. Appl. Log.*, 13(4): 676–719, 2015. doi: 10.1016/j.jal.2015.08.002. URL https://doi.org/10.1016/j.jal.2015.08.002.

[34] M. Eppe and M. Bhatt. A history based approximate epistemic action theory for efficient postdictive reasoning. *J. Appl. Log.*, 13(4):720–769, 2015. doi: 10.1016/j.jal.2015.08.001. URL https://doi.org/10.1016/j.jal.2015.08.001.

[35] GDV. *Compact Accident Research by the German Insurance Association (Unfallforschung der Versicherer)*. 2017.

[36] M. Gebser, R. Kaminski, A. König, and T. Schaub. Advances in gringo series 3. In *LPNMR*, volume 6645 of *Lecture Notes in Computer Science*, pages 345–351. Springer, 2011.

[37] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. *Answer Set Solving in Practice*. Morgan & Claypool Publishers, 2012. ISBN 1608459713, 9781608459711.

[38] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. Clingo = ASP + control: Preliminary report. *CoRR*, abs/1405.3694, 2014.

[39] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[40] S. M. Hazarika. *Qualitative Spatial Change : Space-Time Histories and Continuity*. PhD thesis, The University of Leeds, School of Computing, 2005. Supervisor - Anthony Cohn.

[41] S. M. Hazarika and A. G. Cohn. Abducing qualitative spatio-temporal histories from partial observations. In *KR*, pages 14–25, 2002.

[42] Y. Hou, Z. Ma, C. Liu, and C. C. Loy. Learning lightweight lane detection cnns by self attention distillation. *arXiv preprint arXiv:1908.00821*, 2019.

[43] H. A. Kautz. Reasoning about plans. chapter A Formal Theory of Plan Recognition and Its Implementation, pages 69–124. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991. ISBN 1-55860-137-6. URL http://dl.acm.org/citation.cfm?id=117019.117021.

[44] H. A. Kautz and J. F. Allen. Generalized plan recognition. In T. Kehler, editor, *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, 1986. Volume 1: Science.*, pages 32–37. Morgan Kaufmann, 1986. URL http://www.aaai.org/Library/AAAI/1986/aaai86-006.php.

[45] V. Kondyli and M. Bhatt. Multimodality on the road : Towards evidence-based cognitive modelling of everyday roadside human interactions. In *Advances in Transdisciplinary Engineering :*, volume 11 of *Advances in Transdisciplinary Engineering*, pages 131–142. IOS Press, 2020. ISBN 978-1-64368-104-7. doi: 10.3233/ATDE200018.

[46] V. Kondyli, M. Bhatt, and J. Suchan. Towards a human-centred cognitive model of visuospatial complexity in everyday driving. In S. Rudolph and G. Marreiros, editors, *Proceedings of the 9th European Starting AI Researchers' Symposium 2020 co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago Compostela, Spain, August, 2020*, volume 2655 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2655/paper20.pdf.

[47] R. Kowalski and M. Sergot. *A Logic-Based Calculus of Events*, page 23?51. Springer-Verlag, Berlin, Heidelberg, 1989. ISBN 0387189874.

[48] J. Lee and Y. Wang. A probabilistic extension of the stable model semantics. In *Logical Formalizations of Commonsense Reasoning - Papers from the AAAI Spring Symposium, Technical Report*, volume SS-15-04, pages 96–102. AI Access Foundation, 2015. 2015 AAAI Spring Symposium ; Conference date: 23-03-2015 Through 25-03-2015.

[49] J. Lehner, A. Mitterecker, T. Adler, M. Hofmarcher, B. Nessler, and S. Hochreiter. Patch refinement – localized 3d object detection, 2019.

[50] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer, 2016.

[51] J. Ma, R. Miller, L. Morgenstern, and T. Patkos. An epistemic event calculus for asp-based reasoning about knowledge of the past, present and future. In *LPAR: 19th Intl. Conf. on Logic for Programming, Artificial Intelligence and Reasoning*, volume 26 of *EPiC Series in Computing*, pages 75–87. EasyChair, 2014. doi: 10.29007/zswj.

[52] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00594. URL http://dx.doi.org/10.1109/CVPR.2018.00594.

[53] I. Mani and J. Pustejovsky. *Interpreting Motion - Grounded Representations for Spatial Language*, volume 5 of *Explorations in language and space*. Oxford University Press, 2012. ISBN 978-0-19-960124-0.

[54] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016. URL http://arxiv.org/abs/1603.00831.

[55] R. Miller, L. Morgenstern, and T. Patkos. Reasoning about knowledge and action in an epistemic event calculus. In *COMMONSENSE 2013*, 2013.

[56] P. Muller. A qualitative theory of motion based on spatio-temporal primitives. In A. G. C. et. al., editor, *KR 1998, Italy*. Morgan Kaufmann, 1998.

[57] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang. Spatial as deep: Spatial CNN for traffic scene understanding. In S. A. McIlraith and K. Q. Weinberger, editors, *AAAI 2018*. AAAI Press, 2018.

[58] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao. Mask-guided attention network for occluded pedestrian detection, 2019.

[59] A. Reader and N. Holmes. Examining ecological validity in social interaction: problems of visual fidelity, gaze, and social potential. *Culture and brain*, 4(2):134–146, 2016.

[60] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

[61] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems 2015, Canada*, 2015. URL http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.

[62] T. Schaub and S. Woltran. Special issue on answer set programming. *KI*, 32(2-3):101–103, 2018. doi: 10.1007/s13218-018-0554-8. URL https://doi.org/10.1007/s13218-018-0554-8.

[63] C. P. L. Schultz, M. Bhatt, J. Suchan, and P. A. Walega. Answer set programming modulo ?space-time? In *Proceedings of the International Joint Conference on Rules and Reasoning :*, volume 11092 of *Lecture Notes in Computer Science*, pages 318–326, 2018. doi: 10.1007/978-3-319-99906-7_24.

[64] C. P. L. Schultz, M. Bhatt, J. Suchan, and P. A. Walega. Answer set programming modulo 'space-time'. In C. Benzmüller, F. Ricca, X. Parent, and D. Roman, editors, *Rules and Reasoning - Second International Joint Conference, RuleML+RR 2018, Luxembourg, September 18-21, 2018, Proceedings*, volume 11092 of *Lecture Notes in Computer Science*, pages 318–326. Springer, 2018. doi: 10.1007/978-3-319-99906-7_24. URL https://doi.org/10.1007/978-3-319-99906-7_24.

[65] M. Shanahan. *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press, Cambridge, MA, USA, 1997. ISBN 0262193841.

[66] M. Shanahan. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29(1):103–134, 2005. ISSN 1551-6709. doi: 10.1207/s15516709cog2901_5. URL http://dx.doi.org/10.1207/s15516709cog2901_5.

[67] J. Suchan and M. Bhatt. Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies. In S. Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2633–2639. IJCAI/AAAI Press, 2016. URL http://www.ijcai.org/Abstract/16/374.

[68] J. Suchan and M. Bhatt. The geometry of a scene: On deep semantics for visual perception driven cognitive film,

studies. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9. IEEE Computer Society, 2016. doi: 10.1109/WACV.2016.7477712. URL https://doi.org/10.1109/WACV.2016.7477712.

[69] J. Suchan, M. Bhatt, and P. E. Santos. Perceptual narratives of space and motion for semantic interpretation of visual data. In L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, volume 8926 of *Lecture Notes in Computer Science*, pages 339–354. Springer, 2014. doi: 10.1007/978-3-319-16181-5_24. URL https://doi.org/10.1007/978-3-319-16181-5_24.

[70] J. Suchan, M. Bhatt, and C. P. L. Schultz. Deeply semantic inductive spatio-temporal learning. In J. Cussens and A. Russo, editors, *Proceedings of the 26th International Conference on Inductive Logic Programming (Short papers), London, UK, 2016*, volume 1865, pages 73–80. CEUR-WS.org, 2016.

[71] J. Suchan, M. Bhatt, S. Vardarajan, S. A. Amirshahi, and S. Yu. Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability. *Advances in Cognitive Systems*, 6: 65–84, 2018. ISSN ISSN 2324-8416. URL http://www.cogsys.org/journal.

[72] J. Suchan, M. Bhatt, P. A. Walega, and C. P. L. Schultz. Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. In *32nd AAAI Conference on Artificial Intelligence (AAAI-18), USA*, pages 1965–1972. AAAI Press, 2018.

[73] J. Suchan, M. Bhatt, P. A. Walega, and C. P. L. Schultz. Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects. In S. A. McIlraith and K. Q. Weinberger, editors, *AAAI 2018*. AAAI Press, 2018.

[74] J. Suchan, M. Bhatt, and S. Varadarajan. Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1879–1885. ijcai.org, 2019. doi: 10.24963/ijcai.2019/260. URL https://doi.org/10.24963/ijcai.2019/260.

[75] J. Suchan, M. Bhatt, and S. Varadarajan. Driven by commonsense : On the role of human-centred visual explainability for autonomous vehicles. In *ECAI 2020 :*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2939–2940. IOS Press, 2020. doi: 10.3233/FAIA200463.

[76] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-scnn: Gated shape cnns for semantic segmentation, 2019.

[77] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/tan19a.html.

[78] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 2014. doi: 10.1109/MMUL.2014.29. URL http://dx.doi.org/10.1109/MMUL.2014.29.

[79] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[80] P. A. Walega, M. Bhatt, and C. P. L. Schultz. ASPMT(QS): non-monotonic spatial reasoning with answer set programming modulo theories. In *Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings*, volume 9345 of *Lecture Notes in Computer Science*, pages 488–501. Springer, 2015. doi: 10.1007/978-3-319-23264-5_41.

[81] P. A. Walega, C. P. L. Schultz, and M. Bhatt. Non-monotonic spatial reasoning with answer set programming modulo theories. *Theory Pract. Log. Program.*, 17(2):205–225, 2017. doi: 10.1017/S1471068416000193. URL https://doi.org/10.1017/S1471068416000193.

[82] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[83] Z. Wang and K. Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection, 2019.

[84] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2137, June 2016. doi: 10.1109/CVPR.2016.234.

[85] H. Yu, N. Siddharth, A. Barbu, and J. M. Siskind. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res. (JAIR)*, 52:601–713, 2015. doi: 10.1613/jair.4556.

[86] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation, 2019.

[87] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun. End-to-end interpretable neural motion planner. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[88] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video.

2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6612–6619, 2017.

[89] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu. Class-balanced grouping and sampling for point cloud 3d object detection, 2019.

[90] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. D. Newsam, A. Tao, and B. Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8856–8865. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00906. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_Improving_Semantic_Segmentation_via_Video_Propagation_and_Label_Relaxation_CVPR_2019_paper.html.

[91] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. *Lecture Notes in Computer Science*, page 38–55, 2018. ISSN 1611-3349. doi: 10.1007/978-3-030-01228-1_3. URL http://dx.doi.org/10.1007/978-3-030-01228-1_3.