Self-supervised learning for autonomous vehicles perception: A conciliation between analytical and learning methods

Florent Chiaroni, Mohamed-Cherif Rahal, Nicolas Hueber, Frédéric Dufaux

Self-supervised learning for autonomous vehicles perception: A conciliation between analytical and learning methods

Florent Chiaroni[⋆‡], Mohamed-Cherif Rahal[⋆], Nicolas Hueber[†], and Frédéric Dufaux[‡]

Index Terms

Autonomous vehicle perception, self-supervised learning, semi-supervised learning, scene understanding.

I. Introduction

THE interest for autonomous driving has continuously increased during the last two decades. However, to be adopted, such critical systems need to be safe. Concerning the perception of the ego-vehicle environment, the litterature has investigated two different types of methods. On the one hand, analytical methods, also referred to as hand-crafted, are generally designed from end-to-end. On the other hand, learning methods aim to design their proper representation of the observed scene.

Analytical methods have demonstrated their usefulness for several tasks, including the keypoints detection [1], [2], optical flow, depth map estimation, background subtraction, geometric shape detection, tracking filtering, and simultaneous localization and mapping (SLAM) [3]. Those methods have the advantage to be explainable from end-to-end. However, it is difficult to apply them on high dimensional data for semantic scene analysis. For example, identifying the other users present in an urban scene requires to extract complex patterns from high dimensional data captured by camera sensors.

[⋆]VEDECOM Institute, Department of delegated driving (VEH08), Perception team, {florent.chiaroni, mohamed.rahal}@vedecom.fr

[†]French-German Research Institute of Saint-Louis (ISL), ELSI team, nicolas.hueber@isl.eu

[‡]L2S - CNRS - CentraleSupelec - Univ Paris-Sud - Univ Paris Saclay, {florent.chiaroni, frederic.dufaux}@l2s.centralesupelec.fr

Learning methods are nowadays the most adapted in terms of prediction performances for complex pattern recognition tasks [4] implied in autonomous vehicles scene analysis and understanding. However, the state-of-the-art results are often obtained with large and fully labeled training datasets [5]. Hand-labeling a large dataset for a given specific application has a cost. Another difficulty is to aprehend from end-to-end the learned representations. To overcome the former limitation, transfer learning and weakly supervised learning methods have appeared. Some of them can exploit partially labeled datasets [6], [7], or noisy labeled datasets [8], [9]. Concerning the latter problem, under mild theoretical assumptions on the learning model, we can interpret the predicted outputs. For instance, it is possible to automatically detect the training overfitting [10], to estimate the fraction of mislabeled examples [11], or estimate the uncertainty in the prediction outputs [12].

Another challenge is to **prevent unpredictable events**. Indeed, some scenes unseen during the training can appear frequently in the context of the autonomous vehicle. For instance, an accident on the road can change drastically the appearance and the location of potential obstacles. Thus, even if it is possible to predict when the model does not know what it observes, it may be interesting to confirm it through an analytical process and to adapt the learning model to this novel situation.

It turns out that **self-supervised learning methods (SSL)** have shown in the litterature the ability to address such issues. For instance, the SSL system in [13] won the 2005 DARPA Grand Challenge thanks to its adaptability to changing environments. SSL for autonomous driving vehicles perception is most often based on learning from data which is automatically labeled by an upstream method, similarly to feature learning in [14]. In this paper, we discuss the following aspects of SSL:

- abilities such as online adaptation to the environment evolution, self-supervised evaluation, unnecessity of hand-labeled data, fostering of multimodal techniques [13], and self-improvement. For example, iterative learning reduces progressively the corrupted predictions [15];
- applications enabled by those advantages such as depth map estimation [16], [15], temporal predictions [17], moving obstacles analysis [18], long range vision [13], [19]. For example, the SSL system in [19] learns to extrapolate the appearance of obstacles and traversable areas observable by stereo-vision in a short-range, to identify the long-range obstacles and traversable areas which cannot directly be detected by stereo-vision.

While the cited SSL techniques are respectively designed for a specific use case application,

(a) (b) (c)

Fig. 1. Some self-driving cars. (a) is the self-driving car *Stanley* that won the *DARPA Grand Challenge* using a SSL system equipped with a calibrated monocular camera and a LIDAR sensor [13]. (b) is the autonomous mobile robot *LAGR*. It integrates another SSL vision approach [19] able to identify online the obstacles and road segmentation from a short-range stereovision up to a long-range monocular vision. (c) is the car equipped with the perception sensors used to generate the KITTI dataset [20].

they present some similitudes. In particular, a shared underlying idea is to: *Learn to predict, from a given spatio-temporal information (e.g. a single camera frame [13], [19], [21], [16], [22]), something (e.g. traversable area segmentation [13], [19], depth estimation [16], or moving obstacles segmentation [21], [22]) that can be automatically labeled in another way using additional spatio-temporal information (e.g. stereo-vision camera [19], [16], a temporal sequence [23], or depth sensor [13]).*

We propose to highlight those interdependencies hereafter. In this way, we aim at providing to the reader some analytical, learning and hybrid tools which are transversal to the final application use cases. In addition, the limitations of the presented frameworks are discussed, as well as the perspectives of improvement for self-evaluation, self-improvement, and self-adaptation, in order to address future challenges, as the research actively evolves in the autonomous driving area.

The outline of this article is as follows. After this introduction, we present in Sec. II and III some analytical and learning perception tools relevant to SSL. We follow in Sec. IV by the presentation of existing SSL techniques for some autonomous driving perception applications. Finally we will end by a discussion focusing on limitations and future challenges in Sec. V.

II. Analytical methods

Before the recent growing interest around deep learning methods, many analytical methods (without learning) have been proposed, bringing baseline reference tools for multiple challenging perception tasks in the context of autonomous driving. Some of the most investigated tasks considered in this article are briefly introduced hereafter:

- **Keypoints feature detection**. Before analyzing the sensor data from a relatively high point of view, analytical techniques often require to perform spatial or temporal data matching using **feature detection** methods. More specifically, these methods consist in detecting and extracting local features in the sensor data. These hand-crafted features can be small regions of interest [24]. In order to enable the matching of sensor data, captured from the same scene with different spatial or temporal points of view, such features need to be as invariant as possible to scale, translation, and rotation transformations. The most common sensor data is an image captured by a camera. In this case, competitive feature detectors include SIFT [1], SURF [25], ORB [26]. When a depth sensor is also available, the depth information can be exploited in order to further improve feature detection. For instance, the TRISK method [2] is specifically designed for RGB-D images. More recently, LIDAR has enabled to capture of point clouds. To tackle this new form of sensor data, some feature detection techniques are derived from image ones (e.g. Harris and SIFT). Alternatively, some new approaches such as as ISS [27] are exclusively designed for point clouds. From a practical point of view, implementations of common image feature detectors can be found in image libraries as OpenCV[1], and in point clouds libraries as PCL[2]. Feature detectors are exploited by several autonomous driving perception techniques requiring matching of sensor data, including optical flow, disparity map, visual odometry, SLAM, tracking techniques.
- **Optical flow** is a dense [28] or sparse [29] motion pattern. It can be obtained by computing points or features transformations throughout a temporal images sequence captured from a static or mobile ego-camera point of view. In the context of autonomous driving perception, optical flow can be interesting for background subtraction, motion estimation of the ego-vehicle and surrounding moving obstacles as proposed by Menze et al. [30]. It can also be used, in the absence of additional information, for relative depth map estimation [31] of the surrounding static environment.
- **Depth map estimation** aims at providing image pixels depths, namely the relative or absolute distance between the camera and the captured objects. Several techniques exist to address this task. One of the most common and effective approaches is to compute a disparity map from a stereo-camera. Combined with the extrinsic cameras parameters, such as the baseline distance separating both cameras, the disparity map can be converted into an inversely

[1]https://opencv.org/

[2]http://pointclouds.org/

proportional absolute depth map. Another approach is to project LIDAR points on some of the camera image pixels. It also requires extrinsic spatial and temporal calibrations between both sensors. As mentioned previously, a relative depth map can also be directly deduced on the move from the optical flow obtained with a moving camera in a static scene. Under some assumptions, the absolute depth map can then be obtained, for example with additional accurate GPS and IMU sensors information concerning the absolute pose transformations of the moving camera. The depth map can also be directly obtained with some RGB-D sensors. Depth map is interesting for identifying the 3D shape of objects in the scene. More specifically, in autonomous driving, an absolute depth map is relevant for estimating the distance between the ego-vehicle and detected obstacles. However, we can note that absolute depth map estimation is constraining compared to relative depth map, as at least two jointly calibrated sensors are necessary. Consequently, this has a relative higher financial cost in production. Moreover, extrinsic calibrations can be sensitive to the ego-vehicle physical shocks. Finally such sensor fusions can only offer limited long-range depth estimation, due to fixed baselines with stereo cameras, or sparse point cloud projections with dispersive LIDAR sensors. Nevertheless, relative depth map can be sufficient to detect obstacles and traversable areas. For example, considering the traversable area as a set of planes in the depth map 3D point cloud projection, some template matching techniques can be used [19].

- **Geometric shape detection** techniques such as Hough transform and RANSAC [32] initially aimed at identifying some basic geometric shapes such as lines for lane marking detection, ellipses for traffic lights detection, or planes for road segmentation. In order to deal with sophisticated template matching tasks, techniques such as the hough transform have been generalized (GHT [33]) for arbitrary shape detection. Nonetheless, these techniques require an exact model definition of the shapes to detect. Consequently, they are sensitive to noisy data and are impractical for detection of complex and varying shapes such as obstacles encountered in the context of autonomous driving. Indeed, such objects typically suffer from outdoor illumination changes, background clutter, or non-rigid transformations.
- **Motion tracking** aims at following some data points, features or objects through time. Tracking filters, such as the Extended Kalman Filter (EKF), predict the next motion using the prior motion knowledge. Conversely, objects tracking can be achieved by features or template matching between consecutive video frames. Pixel points and features tracking is interesting for dense or sparse optical flow, as well as visual odometry estimation [34]. Obstacle objects

tracking is very relevant in autonomous driving for modeling or anticipating their trajectories into the ego-vehicle environment. However, on the whole, while some techniques integrate uncertainty, they remain limited when dealing with complex real motion patterns. Pedestrians and drivers behaviour prediction typically requires knowledges about the context. Moreover, mobile obstacles appearence can drastically change depending on their orientation.

- **SLAM techniques.** The complementarity between the above enumerated concepts has been demonstrated through the problem of *simultaneously localizing* the ego-vehicle *and mapping* the surrounding environment (SLAM) [3]. Features matching provides the pose transformations of the moving ego-vehicle. In turn, 3D scaled projections of depth maps combined with the successive estimated poses provide the environment mapping. Tracking filters and template matching can offer some robustness against sensor data noise and drifting localization estimation, as respectively proposed in EKF SLAM [35] and SLAM++ [36] approaches.

To summarize, analytical methods can successfully deal with several perception tasks of significant interest in the context of autonomous driving. In particular, a self-driving vehicle embedding these techniques is able to carry out physical analysis such as the 3D reconstruction modelling of the environment, and dynamic estimations concerning the ego-vehicle and the encountered surrounding mobile obstacles. These techniques have the advantage to be end-to-end explainable in terms of design. This facilitates the identification and prevention of failure modes. However, some critical limitations persist nowadays:

- A lack of landmarks and salient features combined with the presence of dynamic obstacles may entail a severe degradation of the feature detection and matching.
- Severe noisy sensor data induces the same risks.
- It is impossible to achieve dense real-time semantic scene analysis of environments including a wide range of complex shape patterns.

Learning to recognize and predict complex patterns with generalization abilities aims at overcoming such issues, as developped in the next section.

III. Learning methods

Learning methods have demonstrated state-of-the-art prediction performances for semantic tasks during the last two decades. Autonomous driving is a key application which can greatly benefit from these recent developments. For instance, learning methods have been investigated

in this context, for identifying the observed scene context using classification, for detecting the other road users surrounding the ego-vehicle, for delineating the traversable area surface, or for dynamic obstacles tracking.

- **Classification:** It aims at predicting, for a given input sensor sample, an output class label. In order to deal with high dimensional data containing complex patterns, the first stage is generally to extract relevant features using hand-crafted filters or learned feature extractors. For image feature extraction, the state-of-the-art techniques use Convolutional Neural Network (CNN) architectures. The latter are composed of a superposition of consecutive layers of trainable convolutional filters. Then, a second stage is to apply a learning classifier on the feature maps generated as output of these filters. Some commonly used classifiers are the Support Vector Machine (SVM) and the Multi-Layer Perceptron (MLP). Both require a training which is most of the time performed in a fully supervised way on labeled data. The CNN and MLP deep learning models are trained by backpropagating the output prediction error on the trainable weigths up to the input. Concerning the evaluation of these models, a test dataset is required, which is labeled as well. The *Accuracy* metric is commonly used for evaluating the prediction performances, while the F1-Score, an harmonic mean of the precision and recall, is relevant for information retrieval. An image classification application example in autonomous driving is for categorizing the context of the driven road [37].
- **Detection:** It generally identifies in a visual sensor data the regions of interest, which in turn can be classified. A commonly used strategy invariant to scales and rotations applies an image classifier on sliding windows over an image pyramid. Then, several advanced competitive image detection techniques as Faster R-CNN [38], or Yolo [39] have been more recently developped, and have been adapted for road users detection [37].
- **Segmentation:** As its name suggests, this task provides a segmentation of visual sensor data. Three distinct problems can be considered:
  - *Semantic segmentation* assigns a semantic class label to each pixel. An example is road segmentation [37]. State-of-the-art methods generally present a fully convolutional network (FCN) autoencoder (AE) architecture connecting in different ways the encoding part with the decoding part [40]. A standard AE is a generative model composed of an encoder and a decoder learning models which are jointly trained to reconstruct as output the input. In the discussed image segmentation context, it is trained to predict as output a per-pixel classification of the input image pixels.

  - *Instance segmentation* aims at detecting and segmenting each object instance. Examples include foreground segmentation and object detection of potentially moving obstacles [41].
  - *Panoptic segmentation* [4] is a unification of the two previously mentioned segmentation tasks.

  Some models dealing with these segmentation tasks have been adapted for performing per-pixel regression tasks such as dense optical flow estimation [42] or depth map estimation [43].
- **Temporal object tracking** follows the spatial location of selected objects along a temporal data sequence. State-of-the-art learning techniques use variants of the Recurrent Neural Network (RNN) model [44]. Compared to standard filtering techniques, RNNs have the ability to learn complex and relatively long-term temporal patterns in the context of autonomous driving.

While demonstrating competitive prediction performances, the above mentioned learning techniques are fully supervised. In other words, they have in common the limitation to require large-scale fully annotated training datasets. In order to reduce this issue, some other learning strategies have been investigated:

- **Weakly supervised learning:** These techniques can be trained with a partially labeled dataset [6], and eventually with a fraction of corrupted labels [8], [9]. Advantageously, these approaches drastically reduce the need of labeled data.
- **Clustering**: These approaches can be defined as an unlabeled classification stategy, such that it aims at gathering without supervision the data depending on their features similarities. A huge advantage is that no labels are required. However, if it is necessary to associate the clusters obtained with humanly understandable semantic meanings, then a final step of ponctual hand-labeling per-cluster is required. State-of-the-art methods [45] dealing with complex real images mix trainable feature extractors with standard clustering methods such as a Gaussian Mixture Model (GMM) [46].
- **Pre-training:** Some relevant generic visual feature extractors can be obtained by performing a preliminary pre-training of the CNN model on unlabeled or labeled data coming from the target application domain [19] or even from a different one [47].

We note also that in order to aprehend from end-to-end the learned representations, it is possible to identify the training overfitting [10] of deep learning models without validation test supervision.

Furthermore, some learning approaches can estimate the prior of a noisy labeled training dataset [11] or the model uncertainty [12], [48].

Now that some considered analytical and learning methods have been treated separately, the next section shows the complementarity between these two different types of approaches through several Self-Supervised Learning (SSL) systems developped in the context of the autonomous driving vehicle perception.

IV. SSL Autonomous Driving Applications

In the context of autonomous driving applications, we can organize the Self-Supervised Learning (SSL) perception techniques in two main categories:

- High-level scene understanding:
  - road segmentation in order to discriminate the traversable path from obstacles to be avoided
  - dynamic obstacles detection and segmentation
  - obstacles tracking and motion anticipation predictions
- Low-level sensor data analysis, with a particular focus on:
  - dense depth map estimation, which is a potentially relevant input data information for dealing with the previously enumerated scene understanding challenges.

A. Scene understanding

In order to navigate safely, smoothly, or fast when it is required, a self-driving car must perform a path planning adapted to the surrounding environment. The planned trajectories must pass trough traversable areas, while ensuring that surrounding static and dynamic obstacles are avoided. For this purpose, it is necessary to detect and delineate them in advance, but also to anticipate future positions of the mobile ones.

1) Traversable area segmentation: A traversable area can be identified by performing its segmentation over the mapped physical environment. Two different strategies have been successively applied. The former is mainly dedicated to offroad unknown terrain crossing. It entails fully self-supervised training (i.e. without hand-labeled data) systems. The latter, that appeared more recently, is dedicated to urban road analysis. The main difference is that the SSL online systems developped are initialized with a supervised pre-training on hand-labeled data. This preliminary step aims at replacing the lack of landmarks on urban asphalt roads having uniform textures, by prior knowledge.

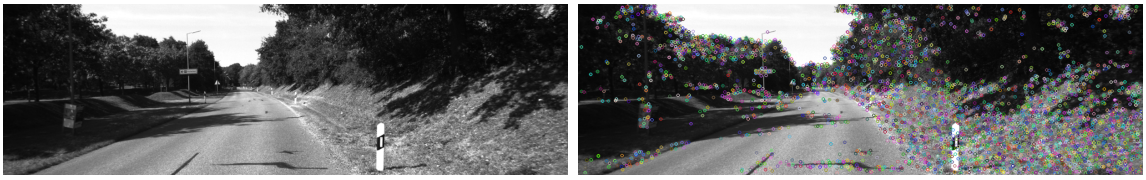
SSL offroad systems: a road segmentation is proposed in [49] by exploiting temporal past information concerning the road appearence on monocular camera images. It considers the close observable area on the current moncular camera frame in front of the car as a traversable road. Next, it propagates optical flow on this area from the current frame up to the past captured frames. Then, it can deduce this close area appearence when it was spatially farther in the past. This past appearence of the actual close traversable area is exploited for producing horizontal line templates using the SSD (sum of squared differences) matching measure. It is combined with a hough transform-based horizon detector to define the image horizontal lines of pixels on which to apply the horizontal 1-D template matching. Next, with the assumption that the actual distant traversable area has roughly the same appearence as the actual close area had in the past, the 1D templates are applied over the current frame to segment the distant traversable area. If the best template matching measure changes abruptly, then it is supposed that the ego-vehicle is going out of the road or that the road appearence has suddenly and drastically changed. The approach in [49] is relevant for providing a long-range road image segmentation using a monocular camera only. However, a major issue is the critical assumption considering the close area as always traversable. If the road aspect changes suddenly, then it is impossible with this SSL strategy to correctly segment this novel road region.

Another SSL road segmentation approach is proposed in [13] dealing naturally with this issue. Instead of using temporal information with the assumption that the close area is always traversable, and in addition to the monocular camera, a LIDAR sensor is used for detecting the obstacles close to the ego-vehicle. Projected on the camera images, LIDAR depth points enable to automatically and sparsely labelize the close traversable area on images pixels. Then, a learning gaussian mixture model (GMM) is trained online to recognize the statistical appearence of these sparse analytically labeled pixels. Next, the learning model is applied on the camera pixels which cannot benefit from the sparse LIDAR points projection, in order to classify them as road pixels or not. In this way, the vehicle can anticipate the far obstacles observable in the monocular camera images, but not in the dispersive LIDAR data. This SSL system enabled the *Stanley* self-driving car, presented in Figure 1(a), to win the *DARPA Grand Challenge*[3] by smoothing the trajectories and increasing the vehicle speed thanks to the anticipation of distant obstacles. This highlighted the interest of combining multiple sensors in a self-driving car.

More recently, with the growing interest for deep learning methods, Hadsell et al. [19] propose to

[3]https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles

use a CNN classifier model instead of the earlier template matching or GMM learning techniques. Moreover, an additional paired camera (i.e. stereo-camera) replaces the LIDAR sensor as in [13]. As offroad terrain traversable areas are not always completely flat, a multi-ground plane segmentation is performed in [19], on the short-range point cloud projection, obtained with the stereo-vision depth map, by using a hough transform plane detector. This technique provides several automatic labels for image patches which are observable in the short-range region. Then, addressing the long-range vision segmentation, the authors firstly train a classifier to predict patches labels automatically estimated within the short-range region. Next, the trained classifier predicts the same labels on the long-range observable image region patches by using a sliding window classification strategy. Concerning the prediction performances, the authors have demonstrated that the online fine tuning of the classifier and the offline pre-taining of its convolutional layers using an unsupervised autoencoder architecture can improve prediction performances. Moreover, an interesting point to note is that instead of using uncertainty or noisy labeled learning techniques, the authors created transition class labels for the boundary image surfaces separating the obstacles from the traversable area. Finally, from an initial 11-12 meters short range stereo-vision, the developped SSL system was able to extrapolate a long-range vision up to 50-100 meters. Nonetheless, in order to estimate the short-range stereo 3D reconstruction, including planar sets of points corresponding to the offroad traversable area, this approach requires the presence of salient visual features in the road regions. This may be impractical for instance on the uniform visual texture of asphalt roads commonly encountered in urban scenarios, as illustrated in Fig. 2.



(a) (b)

Fig. 2. Salient features location on urban ego-vehicle environment. (a) is an arbitrary frame, extracted from the KITTI dataset [20], illustrating an urban asphalt road with the surrounding environment. (b) shows keypoints detected on the left input image using SIFT detector. Keypoints distribution is dense on the offroad region observable in the image right side, and sparse on the asphalt road observable in the image center.

Pre-trained SSL urban road systems: Some other online SSL techniques deal with this issue by exploiting a classifer pre-trained offline on hand-labeled data [50], [51].

The automatic labeling step previously performed with analytical methods is replaced in [50] by

an SVM classifier pre-trained offline using a human annotated dataset. In this way, this approach is also compatible with uniform asphalt road surfaces. However, compared to the previously presented SSL offroad approaches, it requires hand-labeled data.

A hybrid path segmentation technique is proposed in [51]. It combines a 3D traversability cost map obtained by stereo-vision, and an SVM classifier pre-trained offline over a human annotated dataset. Six different ground surfaces are considered to train the classifier: asphalt, big gravel, small gravel, soil, grass, bushes and stones. The strategy is as follows. SVM predictions refine online the cost map concerning the flat regions. In turn, the 3D traversability cost map obtained without supervision is exploited to update online some mis-classifications of the pre-trained classifier.

To sum up regarding these road segmentation SSL systems, we can notice that while the sensor data and the analytical and learning models are different for each approach, the online process remains essentially the same. The first stage always consists in generating automatic labels by using additional temporal [49], sensor [13], [19], or prior knowledge information [50], [51]. Then, a second stage trains or updates online a classifier, such that it can be used to provide a long-range or refine road segmentation. Overall, while the short-range visions based on depth sensors aims at ensuring the reliable detection of close obstacles, using such SSL vision techniques in static environments directly enables to anticipate the path planning evolution. Consequently, it is possible to increase the maximum speed velocity of the self-driving car [13], while preserving smooth trajectories [19].

Now that we have presented some SSL techniques dealing with limited depth sensors in static environments, we focus on dynamic obstacles, as they represent the other potential road users interacting with the ego-vehicle in the shared surrounding environment.

2) Dynamic obstacles analysis: We start by presenting an SSL approach [21] based on a binary per-pixel segmentation of dynamic obstacles. Then, we present its extension [18] for dynamic obstacles instance segmentation, such that the different road users can be separated.

SSL for dynamic obstacles per-pixel segmentation: a per-pixel binary segmentation of dynamic obstacles is proposed in [21], using temporal image sequences captured with a monocular camera installed on a mobile urban vehicle. The approach firstly separates sparse dynamic keypoints features from the static ones, by applying a RANSAC technique over the optical flow between consecutive frames. Then, the automatically produced per-pixel dynamic labels are transferred as input of a learning Gaussian Process (GP) model. Next, the learned model

extrapolates this knowledge to label as dynamic the pixels following the same visual propoerties than the ones previously automatically identified as dynamic. The whole process is achieved during an online procedure. The system is evaluated on a hand-labeled dataset. This SSL strategy has the advantage to provide the background subtraction from a moving camera, while extrapolating a dense per-pixel segmentation of the dynamic obstacles from sparse detected keypoints only. However, this technique cannot provide per-obstacles analysis as it merely predicts a binary mask of pixels corresponding to dynamic obstacles.

The technique in [18] extends the previous approach for SSL multi-instance segmentation by using temporal image sequences captured with a monocular camera installed on a mobile urban vehicle. The authors apply, over the mobile keypoints detected by [21], a clustering method using the tracked keypoints information such as their spatial location and motion pattern features. The multi-instance segmentation of dynamic obstacles is evaluated on a hand-labeled video sequence of the KITTI dataset [20].

Overall, the authors state that some issues shared with analytical methods persist in their approache. If the dynamic obstacles shadows are projected on the background, then the latter are considered as dynamic as well. Moreover, the segmentation of distant dynamic obstacles can be missed if the corresponding keypoints variations are considered as noise due to the difficulty to detect the corresponding slight optical flow variations. Furthermore, if a dynamic obstacle, large or close to the sensor, represents the majority of the image keypoints, then this given obstacle is likely to be treated as the static background scene.

Nonetheless, it is important to bear in mind that these approaches present state-of-the-art competitive performances for dynamic obstacles detection and segmentation without training or pre-training on annotated data. In addition, the method in [18] provides interesting tools to analyze on the move the dynamic obstacles, for example to separately track them and learn to predict their intention.

The next focus is on SSL techniques designed for object tracking and temporal predictions in urban road scene evolution, including dynamic obstacles.

3) Temporal tracking predictions: In order to deal with object appearence changes, a competitive SSL tracking technique [52] proposes an online adaptive strategy combining tracking, learning, and object detector real-time modules. However, in the context of autonomous driving, it may be often necessary to simultaneously track, and even anticipate the trajectories of several surrounding road users. Moreover, being able to consider the interactions between each road user requires

under particular circumstances some complex motion pattern analysis.

It turns out that some SSL approaches propose to deal with this challenge by focusing the prediction effort on the entire scene in a unified way, rather the on every obstacles independently. The SSL *deep tracking* system [23][4] learns to predict the future state of a 2D LIDAR occupancy grid. This is achieved by training an RNN on the latent space of a CNN autoencoder (AE) which is applied on the input occupancy grid considered as an image. Each cell of the grid is represented by a pixel, which can be color-coded as occluded, void, or as an obstacle surface. Consequently, the model can be trained from end-to-end by learning to predict the next occupancy grid states using the past and current grid states. Solely the prediction output error of non occluded cells is backpropagated during the training. By definition, this system can perform a self-evaluation by computing a per-pixel photometric error between the predicted occupancy grid and the real future observed occupancy grid at the same temporal instant. This technique has the advantage of being compatible with complex motion patterns compared to Bayesian and Kalman tracking techniques. In addition, the training process enables to predict the obstacles trajectories even during occlusions. The major interest of *deep tracking* is that, as the model learns to predict a complete scene, it naturally considers interactions between each dynamic obstacle present in the scene. In [17], the *deep tracking* model is extended for a real mobile LIDAR sensor by adding a spatial transformer module in order to take into consideration the displacements of the ego-vehicle with respect to its environment during objects tracking.

In turn, these tracking approaches provide the tools to collect motion pattern information of surronding dynamic obstacles such that this information may help to classify obstacles depending on their dynamic properties [53].

B. Low-level sensor data analysis

We address now the sensor data analysis for low-level information estimation in the context of autonomous driving. Compared to the previous methods, the attention has mainly focused recently on SSL depth map estimation from monocular or stereo cameras.

1) SSL Depth map estimation: The self-supervised depth map estimation approach presented in [16] predicts a depth map from a monocular camera without relying on annotated depth maps. The pose transformation between both left and right cameras is known. The SSL strategy is as follows.

[4]Such an approach could be categorized as unsupervised. However, we make the choice in this article to consider that exploiting during the training an additional future temporal information, not availbable during the prediction step, is a type of self-supervision.

First, the left camera frame is provided as input to a fully CNN model trained from scratch to predict, the corresponding depth map. Second, an inverse warping is performed by combining the predicted left depth map with the right camera frame in order to output a synthesized frame similar to the input left frame. In this way, an SSL photometric reconstruction error can be computed in output of the decoder part. Next, this per-pixel error is directly used to train the encoder weights using an SGD optimization technique. While not requiring pre-training, nor annotated ground-truth depths, this approach presents prediction preformances comparable with the state-of-the-art fully supervised monocular techniques. However, the ground truth pose transformation, related to the inter-view displacement between both cameras, is required.

Following a similar idea, another technique is proposed in [15]. It is trained to reconstruct, from a given frame, the second frame taken from a different point of view. It generates a depth map using a stereo camera during the training step, but also during the prediction step. This makes the approach more robust, such that it becomes competitive with standard stereo matching techniques. Moreover, the constraint of preserving two cameras and the pose transformation ground truth for predictions, enables in counterpart to perform online learning. This may be interesting for dealing with novel ego-vehicle environments unseen during the training.

In order to overcome the necessity of the pose transformation ground-truth, Zhou et al. [54] propose to predict, from a temporal sequence of frames, the depth map with a learning model, and the successive camera pose transformations with another learning model. Both models are trained together from end-to-end for making the novel view synthesis of the next frame. However, such a pose transformation estimation implies that the predicted depth map is defined up to a scale factor.

A more modular technique [47] exploits either temporal monocular sequences of frames as in [54], the paired frames of a stereo camera as in [15], or to jointly exploit both temporal and stereo information. This framework also deals with the false depth estimation of moving obstacles by ignoring, during training, the pixels not varying between two consecutive temporal frames. It also deals with occluded pixels when the captured point of view changes by using a minimum reprojection loss.

To summarize, low-level analysis techniques for depth map estimation have demonstrated that SSL strategies without using ground truth labels can bring state-of-the-art solutions competitive with fully supervised techniques.

Overall, the SSL techniques presented in this section support the following conclusion. By

exploiting the complementarity between analytical and learning methods, it is possible to address several low-level and challenging autonomous driving perception tasks, without necessarily requiring a fully annotated dataset.

The next section presents self-supervised learning limitations and future challenges for autonomous driving perception applications.

V. Limitations and future challenges

In the context of autonomous driving, some limitations remain in the presented SSL perception systems and open future research perspectives:

- *Catastrophic forgetting:* During the online learning procedure, the trainable weights of the model may require unnecessary repetitive updates for detecting a given pattern throughout the environment exploration. In fact, when a learning model is continuously specialized for dealing with the latest data, the likelihood increases that the model simultaneously forget the potentially relevant formerly learned patterns. It turns out that it is possible to deal with this *catastrophic forgetting* issue when using neural networks [55]. For future research directions, it may be interesting to combine such incremental learning techniques with the presented SSL frameworks.
- Concerning the scene depth map estimation solely based on temporal analysis:
  - the presence of dynamic obstacles in the scene during the learning stage can result in poor estimates of the observed scene. As discussed in [21], further research on SSL for potentially dynamic obstacles delineations on the sensor data may help to deal with this issue.
  - the current state-of-the-art techniques cannot estimate the real depth map without requiring a supervised scaling factor. The latter is generally obtained by estimating the real metric values of the pose transformation between two consecutive camera viewpoints. As proposed in the supervised detector *Deep MANTA* [56], it may be interesting to recover automatically this scale factor by using some template matching techniques on the observable objects of the scene.
- Concerning the online self-evaluation, some of the presented systems require a baseline reference obtained analytically [19]. However, if we consider that the analytical processes, considered as ground-truth labeling techniques, are likely to generate some noisy labels, it may be interesting to investigate some future reasearch on how to evaluate this prior noise from the learning model point of view [11], and how to deal with it [9].

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] M. Karpushin, G. Valenzise, and F. Dufaux, "Keypoint detection in rgbd images based on an anisotropic scale space," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1762–1771, 2016.

[3] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[4] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollr, "Panoptic segmentation," *arXiv preprint arXiv:1801.00868*, 2018.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[6] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama, "Theoretical comparisons of positive-unlabeled learning against positive-negative learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 1199–1207.

[7] F. Chiaroni, M.-C. Rahal, N. Hueber, and F. Dufaux, "Learning with a generative adversarial network from a positive unlabeled dataset for image classification," in *IEEE International Conference on Image Processing*, 2018.

[8] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. Wijewickrema, and J. Bailey, "Dimensionality-driven learning with noisy labels," *arXiv preprint arXiv:1806.02612*, 2018.

[9] F. Chiaroni, M. C. Rahal, N. Hueber, and F. Dufaux, "Hallucinating a Cleanly Labeled Augmented Dataset from a Noisy Labeled Dataset Using GANs," in *IEEE International Conference on Image Processing*, 2019.

[10] M. E. Houle, "Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications," in *International Conference on Similarity Search and Applications*. Springer, 2017, pp. 64–79.

[11] S. Jain, M. White, and P. Radivojac, "Estimating the class prior and posterior from noisy positives and unlabeled data," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 2693–2701.

[12] Y. Gal, "Uncertainty in Deep Learning," Ph.D. dissertation, PhD thesis, University of Cambridge, 2016.

[13] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain." in *Robotics: science and systems*, vol. 38. Philadelphia, 2006.

[14] L. Jing and Y. Tian, "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey," *arXiv preprint arXiv:1902.06162*, 2019.

[15] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[16] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[17] J. Dequaire, P. Ondruska, D. Rao, D. Wang, and I. Posner, "Deep tracking in the wild: End-to-end tracking using recurrent neural networks," *The International Journal of Robotics Research*, p. 0278364917710543, 2017.

[18] A. Bewley, V. Guizilini, F. Ramos, and B. Upcroft, "Online self-supervised multi-instance segmentation of dynamic objects," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1296–1303. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6907020/

[19] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[21] V. Guizilini and F. Ramos, "Online self-supervised segmentation of dynamic objects," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4720–4727.

[22] D. Pathak, R. Girshick, P. Dollr, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2701–2710.

[23] P. Ondruska and I. Posner, "Deep tracking: Seeing beyond seeing using recurrent neural networks," *arXiv preprint arXiv:1602.00991*, 2016.

[24] C. G. Harris, M. Stephens *et al.*, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf." in *ICCV*, vol. 11, no. 1. Citeseer, 2011, p. 2.

[27] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 2009, pp. 689–696.

[28] G. Farnebck, "Two-frame motion estimation based on polynomial expansion," *Image analysis*, pp. 363–370, 2003.

[29] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[30] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.

[31] K. Prazdny, "Egomotion and relative depth map from optical flow," *Biological cybernetics*, vol. 36, no. 2, pp. 87–102, 1980.

[32] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[33] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[34] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE robotics & automation magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[35] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1052–1067, 2007.

[36] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[37] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving," *arXiv preprint arXiv:1612.07695*, 2016. [Online]. Available: https://arxiv.org/abs/1612.07695

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[41] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[42] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[43] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, Oct 2016.

[44] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[45] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[46] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[47] C. Godard, O. Mac Aodha, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv:1806.01260*, 2018.

[48] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015. [Online]. Available: https://arxiv.org/abs/1511.02680

[49] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive Road Following using Self-Supervised Learning and Reverse Optical Flow." in *Robotics: science and systems*, 2005, pp. 273–280.

[50] S. Zhou, J. Gong, G. Xiong, H. Chen, and K. Iagnemma, "Road detection using support vector machine based on online learning and evaluation," in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 256–261.

[51] H. Roncancio, M. Becker, A. Broggi, and S. Cattani, "Traversability analysis using terrain mapping and online-trained terrain type classifier," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 1239–1244.

[52] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, July 2012.

[53] M. Fathollahi and R. Kasturi, "Autonomous driving challenge: To Infer the property of a dynamic object based on its motion pattern using recurrent neural network," *arXiv preprint arXiv:1609.00361*, 2016.

[54] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[55] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, p. 201611835, 2017.

[56] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep manta: A coarse-to-fine many-task

network for joint 2d and 3d vehicle analysis from monocular image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.