Computer Vision and Image Understanding 184 (2019) 1–8

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Automotive radar and camera fusion using Generative Adversarial Networks

Vladimir Lekic [a,*], Zdenka Babic [b]

[a] Daimler AG - Group Research & Mercedes-Benz Cars Development, Ulm, Germany
[b] Faculty of Electrical Engineering, University of Banja Luka, Bosnia and Herzegovina

ARTICLE INFO

Communicated by N. Paragios

Keywords:
Radar
Camera
Unsupervised learning
Generative Adversarial Networks
Driver assistance
Highly automated driving
Image processing
Computer vision

ABSTRACT

Radar sensors are considered to be very robust under harsh weather and poor lighting conditions. Largely owing to this reputation, they have found broad application in driver assistance and highly automated driving systems. However, radar sensors have considerably lower precision than cameras. Low sensor precision causes ambiguity in the human interpretation of the measurement data and makes the data labeling process difficult and expensive. On the other hand, without a large amount of high-quality labeled training data, it is difficult, if not impossible, to ensure that the supervised machine learning models can predict, classify, or otherwise analyze the phenomenon of interest with the required accuracy. This paper presents a method for fusing the radar sensor measurements with the camera images. A proposed fully-unsupervised machine learning algorithm converts the radar sensor data to artificial, camera-like, environmental images. Through such data fusion, the algorithm produces more consistent, accurate, and useful information than that provided solely by the radar or the camera. The essential point of the work is the proposal of a novel Conditional Multi-Generator Generative Adversarial Network (CMGGAN) that, being conditioned on the radar sensor measurements, can produce visually appealing images that qualitatively and quantitatively contain all environment features detected by the radar sensor.

1. Introduction

Radar (Radio Detection and Ranging) sensors and cameras are a crucial part of the sensor setup of the driver assistance and the highly automated driving systems. A sensor in such systems is required to maintain its precision and robustness under different, often adverse, system environment conditions. Precision describes the sensor's ability to reproduce the measurements, while robustness requires that a sensor maintain its measurement accuracy. Most sensors are particularly suited to fulfilling either one or the other of these requirements, but none is ultimately able to fulfill both. For example, camera systems produce precise images of the environment but are very sensitive under, among other things, poor lighting conditions. Radar sensors, on the other hand, fulfill the robustness requirement but have considerably lower precision than cameras.

Radar is used in today's driver assistance systems to determine range, velocity, azimuth angle, and elevation angle of objects in the vehicle surroundings. Relying on the Doppler effect, a radar has an inherent ability to accurately measure the relative velocity of the detected objects, and therefore to easily discriminate between the dynamic and the static objects it detects. Dynamic objects, like pedestrians, cyclists or moving vehicles, are usually stored in a so-called dynamic-object-list and tracked over time using some of the well-known and proven probabilistic tracking algorithms (Cho et al., 2014), and therewith not in the focus of this paper. Static road objects, like road boundaries, bridges, tunnels or parked vehicles are usually extending over a large area, and due to the low precision of an automotive radar, the measurements belonging to these object are difficult to cluster and classify. A common approach to increase the measurement precision is to integrate or fuse the static measurements over time, using a tessellated representation of the sensor environment called evidence grid (Pagac et al., 1996). An example of such evidence grid fusion, with the corresponding camera image of the scene is shown in Fig. 1. The belief in the occupancy state of an area, that the cell in a grid represents, is inferred from the radar measurements, and combined over time into free (Fig. 1(b)) and occupancy (Fig. 1(c)) grid layers using the Dempster–Shafer combination rule (Wu et al., 2002). As the color-map in Fig. 1(d) shows, beliefs in the occupancy state of the cells range from 0 to 1. If there is no evidence about the occupancy state of a cell, the belief is 0. This occurs, for example, during the initialization phase, when there are no measurements presented to the model, and the state of the environment is unknown. If there is strong evidence about the occupancy state, and such evidence is confirmed over multiple measurement cycles, the belief converges to 1. The potential for the application of supervised machine algorithms on such grids is, of course, recognized in the scientific community (Dubé et al., 2014; Lombacher et al., 2016). Until now, these algorithms have

* Corresponding author.
E-mail addresses: vladimir.lekic@daimler.com (V. Lekic), zdenka.babic@etf.unibl.org (Z. Babic).

https://doi.org/10.1016/j.cviu.2019.04.002
Received 27 July 2018; Received in revised form 3 April 2019; Accepted 7 April 2019
Available online 24 April 2019

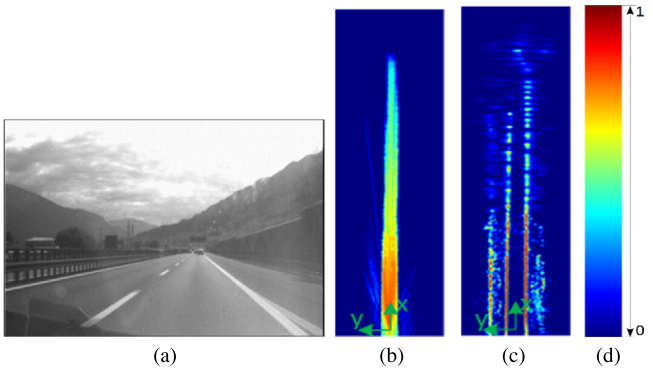


Fig. 1. Multi-modal scene recordings: (a) camera image of the scene, (b) free and (c) occupancy grid layers, and (d) color-map used for the visualization of the evidence grid layers. It should be noted that the moving vehicle on the road, visible on the camera image, is not present in the free and occupancy grid layers, as these contain only the static radar measurements.

been trained using the human-labeled grid data-sets. To be able to label the grid data correctly, human labelers need to perform cross-modality matching of the grid objects with the corresponding objects in the scene images, recorded by an additional camera system. Often, in the attempt to find the correspondence between the objects, ambiguity arises in even relatively simple road scenes, making the labeling tasks challenging and costly. This presents a significant bottleneck for the successful application of the supervised machine learning algorithms for classification of the radar measurements. On the other hand, camera images are practically irreplaceable for object detection and image classification tasks. With the popularization of end-to-end convolutional neural networks for semantic image segmentation (Long et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2018), several labeled image data-sets targeting highly automated driving systems have even become publicly available (Cordts et al., 2016; Neuhold et al., 2017).

The primary hypothesis in this work is that through data fusion of the radar sensor measurements and the camera images it is possible to produce more consistent, accurate, and useful information than that provided by each of the sensors independently. We differentiate between two general data fusion strategies, namely: feature-fusion and semantic-fusion. Feature-fusion refers to a process of fusing data from different modalities in a feature space, for example combining two feature vectors to obtain a single feature vector. Semantic-fusion, on the other hand, refers to a process of fusing data in a semantic space, for example combining pixel semantic classes from two or more data sources. To enable feature-fusion of multi-modal scene recordings we proposed a fully-unsupervised machine learning algorithm based on the Conditional Multi-Generator Generative Adversarial Networks (CMGGANs). During an adversarial training, if conditioned on the appropriate radar sensor data, generators of CMGGANs are able to discover multiple, but disjointed modes of real data distribution. By sampling from the mixture of all learned conditional probability distributions, the algorithm produces the images that contain all the environment features detected by the radar sensor. In a semantic-fusion stage, these generated images can be fused at a semantic level with original camera images of the same scene to increase the robustness of a camera, or their semantic labels can be used to enable an efficient end-to-end semantic segmentation of the radar data. CMGGANs are described in Section 2. Section 3 describes the radar–camera fusion principles based on CMGGANs. The experimental setup and the obtained results are presented in Section 4, while the concluding remarks are given in Section 5.

2. Conditional multi-generator GANs

2.1. Introduction to GANs

Generative adversarial networks (GANs) (Goodfellow et al., 2014) were introduced as an alternative framework for training generative models to overcome the difficulty of approximating often intractable data generating distributions. A high-level architecture of a GAN is shown in Fig. 2. GANs consist of the two adversarial models playing a zero-sum minimax game:

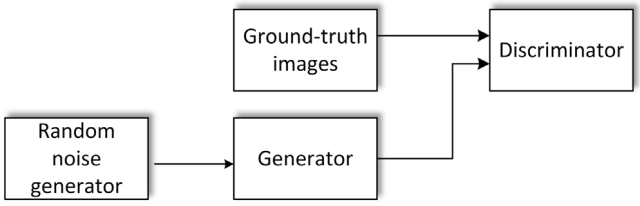


Fig. 2. GAN architecture.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x,\theta_d))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z,\theta_g)))]. \tag{1}$$

A generative model $G(z,\theta_g)$ captures the data distribution $p_g(x)$ over the data x. It basically represents a mapping from a prior $p_z(z)$, defined over the input noise variables z, to the data distribution $p_g(x)$. A discriminative model $D(x,\theta_d)$ is trained to estimate the probability a sample x is drawn from the training set distribution $p_{data}(x)$ rather than from the $p_g(x)$. Variables θ_g and θ_d represent the trainable model parameters.

Although it has been shown in various works (Radford et al., 2015; Pan et al., 2017) that GANs can generate high-quality images while capturing semantic attributes of the training images, they still suffer from the mode-collapse problem (Goodfellow, 2017). As a consequence of this problem, a generator can learn to produce only a narrow variety of modes. Many attempts have been made to improve GANs in this sense. Mirza and Osindero (2014) in their Conditional GANs condition the generator and discriminator models on additional information to direct the data generation process. Chen et al. (2016) propose an interesting InfoGAN architecture that, in addition to the original GAN setup, also contains a regularization parameter that maximizes the mutual information between a small subset of the input latent variables:

$$\min_{G,Q} \max_D V_{InfoGAN}(G,Q,D) = V(D,G) - \lambda L_I(G,Q), \tag{2}$$

where $V(D,G)$ is the value function defined in Eq. (1), $L_I(G,Q)$ is a variational lower bound of the mutual information, and λ is a hyper-parameter. Further, to improve the generation diversity, Hoang et al. (2017) propose a Mixture GAN (MGAN) architecture. The minimax game for K generators G_k, a classifier C, and a discriminator D is formulated as:

$$\min_{G_{1:K},C} \max_D V_{MGAN}(G_{1:K},C,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_{model}(x)}[\log(1-D(x))] - \beta\left\{\sum_{k=1}^{K}\pi_k \mathbb{E}_{x\sim p_{G_k}(x)}[\log C(x)]\right\}, \tag{3}$$

where $p_{model}(x) = \sum_{k=1}^{K}\pi_k p_{G_k}(x)$ represents the mixture of the generator distributions, $C(x)$ is the probability that sample x is generated by generator G_k, and $\beta > 0$ is the diversity hyper-parameter. In other words, in addition to the goal that the mixture of all distributions $p_{model}(x)$ approximates the real data distribution $p_{data}(x)$, at equilibrium generators are also required to cover different modes of the data distribution. Although the authors partially remedy the mode-collapse problem in simple synthetic and real-world data-sets, this architecture still cannot direct the generators to discover the modes of interest in the real data distributions.

2.2. Proposed model

Similarly to the InfoGANs (Chen et al., 2016), we decomposed the input vector into two parts: incompressible noise z and the structured latent variables y_k, where $k \in \{1, \dots, K\}$. Our intention was to use the domain-specific knowledge to introduce the set of meaningful latent variables that will target structured semantic information of the data, and which would force the model to generalize better. Each of the K generators G_k is conditioned on the input variable y_k and the minimax game is defined as:

$$\begin{aligned}&\min_{G,C}\max_{D} V^{(K)}_{CMGGAN}(G,C,D) = \mathbb{E}_{x\sim p_{data}}[\log D(x)]\\&+ \mathbb{E}_{x\sim p_{model}}[\log(1-D(x))]\\&- \beta\sum_{k}\sum_{l\neq k}\pi_k\Big(\mathbb{E}_{x\sim p_{G_k}}[\log C_{k,l}(x|y_k)]\\&+ \mathbb{E}_{x\sim p_{G_l}}[\log(1-C_{k,l}(x|y_k))]\Big).\end{aligned} \tag{4}$$

Each pair of generators G_k and G_l, where $k \neq l$, has an assigned classifier $C_{k,l}$. There are $n(n-1)/2$ such classifiers, and analog to the generator model, each classifier $C_{k,l}$ is also conditioned on the input variable y_k. The idea behind the binary classifiers $C_{k,l}$ is to guide the generators G_k and G_l to discover different modes of the real data distribution that are related to the image features represented by the conditional latent variables. At the same time, discriminator D forces the generators to generate a mixture distribution that represents the real data distribution well. Without the classifiers, the network architecture is basically the same as K independent Conditional GANs. In such an architecture the training of any one generator is independent of the training of the rest of $K-1$ generators, and prone to all aforementioned problems of a single generator training. In that sense, the role of CMGGAN classifiers could be seen as bringing the cross-generator-dependency into the generator training procedure. Moreover, through such dependency and by conditioning the classifiers on the input variables, generators are also enforced to learn the different data distribution modes.

3. Radar-camera fusion using CMGGANs

It is assumed that the radar sensor readings are modeled using an evidential, grid-based representation of spatial information. As already mentioned in Section 1, in such a model, each grid cell represents a belief in the occupancy states: free $p_{free}(x_1, x_2)$ and occupied $p_{occ}(x_1, x_2)$. A pair (x_1, x_2) represents the measurement position within the coordinate system with the origin at the sensor's mounting position. We refer to the entire collection of the occupancy state estimates as a grid layer, namely free grid layer and occupancy grid layer. By incorporating these layers, together with camera images into the CMGGAN network, in a fully unsupervised manner, we enable a feature-fusion of the radar and camera data. The high-level architecture of such a network is shown in Fig. 3. As seen in that figure, free and occupancy grid layers, further on represented by the variables y_1 and y_2 respectively, are used to condition two out of the three generators in the model. The third generator was not conditioned. Following these assumptions, Eq. (4) can be rewritten as:

$$\begin{aligned}&\min_{G,C}\max_{D} V^{(3)}_{CMGGAN}(G,C,D) = \mathbb{E}_{x\sim p_{data}}[\log D(x)]\\&+ \mathbb{E}_{x\sim p_{model}}[\log(1-D(x))] - \beta\Big[\pi_1\big(\mathbb{E}_{x\sim p_{G_1}}[\log C_{1,2}(x|y_1)]\\&+ \mathbb{E}_{x\sim p_{G_2}}[\log(1-C_{1,2}(x|y_1))]\big) - \pi_2\big(\mathbb{E}_{x\sim p_{G_2}}[\log C_{2,3}(x|y_2)]\\&+ \mathbb{E}_{x\sim p_{G_3}}[\log(1-C_{2,3}(x|y_2))]\big) - \pi_3\big(\mathbb{E}_{x\sim p_{G_3}}[\log C_{3,1}(x)]\\&+ \mathbb{E}_{x\sim p_{G_1}}[\log(1-C_{3,1}(x))]\big)\Big].\end{aligned} \tag{5}$$

Analogously to the discussion in Section 2, the intuition behind conditioning the generators and the classifiers with the radar measurements is: to guide the generator conditioned on the free grid layer in learning how to generate the features that are on the image considered to be a free drivable space, and to guide the generator conditioned on the occupancy grid layers to learn how to generate the features that represent the static obstacles detected by radar. Since the radar has a limited range and accuracy, the image features not detected by the radar should be generated completely randomly by the third generator, which has only random noise as input.

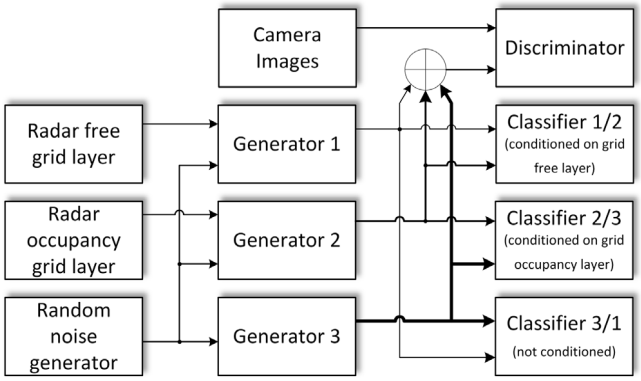
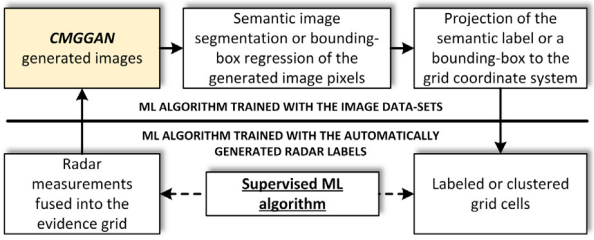
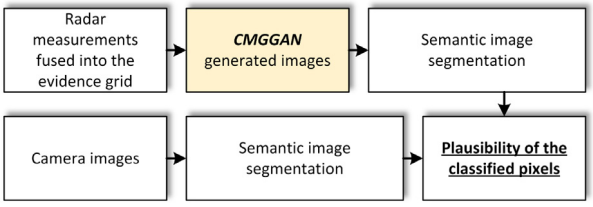


Fig. 3. High-level CMGGAN architecture for feature-fusion of radar and camera data.



(a) Semantic segmentation of the radar data



(b) Plausibilization of the camera image pixel classifications

Fig. 4. Semantic-fusion principles based on the artificially generated camera images.

In addition to the described feature-fusion of radar measurements and camera images, two semantic-fusion principles, based on artificially generated camera images are shown in Fig. 4. As already mentioned in the introductory section, labeling raw radar data is a difficult process, from the labeling procedure and consequently from the commercial point of view. Fig. 4(a) shows an experimental setup which enables an end-to-end training of a supervised machine learning algorithm for semantic segmentation or object bounding-box regression of a pure radar data, completely avoiding the radar data labeling process. Furthermore, popular deep-learning algorithms for semantic image segmentation calculate a categorical label for each image pixel, without providing any measure of uncertainty for these calculations. With the label probabilities, on the other hand, it would be possible to counterfeit phenomena such as adversarial examples (Szegedy et al., 2013), or to increase the robustness property of an algorithm by filtering the labels over time using some of the well-established statistical methods. An approach to assigning the probability measure to the pixel labels using the radar measurements is shown in Fig. 4(b).

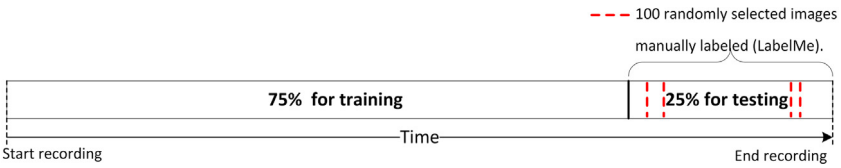


Fig. 5. Structure of the data-set.

To establish confidence in the soundness and usefulness of the proposed CMGGAN model for these use-cases, we designed validation procedure that performs both, subjective and objective model assessment, as it will be shown in the following section.

4. Experiments and analysis

4.1. Experiment setup

The experimental data-set consists of 9,850 synchronized pairs of fused radar sensor measurements and camera images, recorded during one continuous drive on a highway. One such pair is shown in Fig. 1, and the overall data-set structure is shown in Fig. 5. 75% of the data-set is selected as the training data-set, while the rest is used for the model testing and validation. From the test data-set, 100 camera images were randomly selected and manually labeled. Image labeling was performed using the LabelMe tool (Russell et al., 2008), and the labelers had the task of drawing a polygon around the pixels that represent the drivable free space around the system-vehicle. These labeled images, together with the ground-truth camera images and the radar counterparts represent the validation data-set.

The experimental model consists of 3 generators, 3 classifiers and 1 discriminator network, as described in Section 3. All the networks are designed following the principles given in Radford et al. (2015). Detailed architectures of these networks are shown in Fig. 6. For the generator models, each noisy input is extended with the corresponding conditional variable. Discriminator and classifier models have the same architecture of the layers, except that the respective conditional input variables are introduced at the last hidden layer of the classifiers. All inputs to the networks are linearly scaled to have zero mean and unit norm.

4.2. Model evaluation

For the generative models, the performance of the model is often measured based on the quality assessment of the generated images. Usually, this is done by having human annotators judge the visual quality of samples (Denton et al., 2015), or alternatively, by calculating the Inception Score (Salimans et al., 2016), Fréchet Inception Distance (Heusel et al., 2017), or some other objective image quality measure. The validation procedure is presented in Fig. 7. Performance of CMGGANs is compared to the already mentioned InfoGANs (Chen et al., 2016) and Conditional GANs (CGANs) (Mirza and Osindero, 2014). Training of the proposed network and the reference networks is performed over $N = 150$ training epochs. Hyper-parameter sweep is performed over the learning rate $\eta \in \{0.0001, 0.00001\}$ for all the networks, as well as over the mutual information penalty $\lambda \in \{0.5, 1.0\}$ for InfoGAN, and gradient penalty $\gamma \in \{0.5, 1.0\}$ for both, InfoGAN and CGAN, as they are both trained using the Wasserstein loss function (Arjovsky et al., 2017).

The Inception Score is a popular method for measuring the quality of generated image samples. The authors proposed applying an Inception-v3 (Szegedy et al., 2015) network, pre-trained on ImageNet (Deng et al., 2009), to predict the class labels for the generated images, and then calculating the KL-divergence between the conditional label distribution and the marginal label distribution. One shortcoming of the Inception Score is that it evaluates the generated samples "in a vacuum", comparing the statistics only between the generated samples. To overcome this shortcoming, Heusel et al. (2017) proposed the Fréchet Inception Distance (FID), that improves on the Inception Score by actually comparing the statistics of generated image samples to ground-truth images.

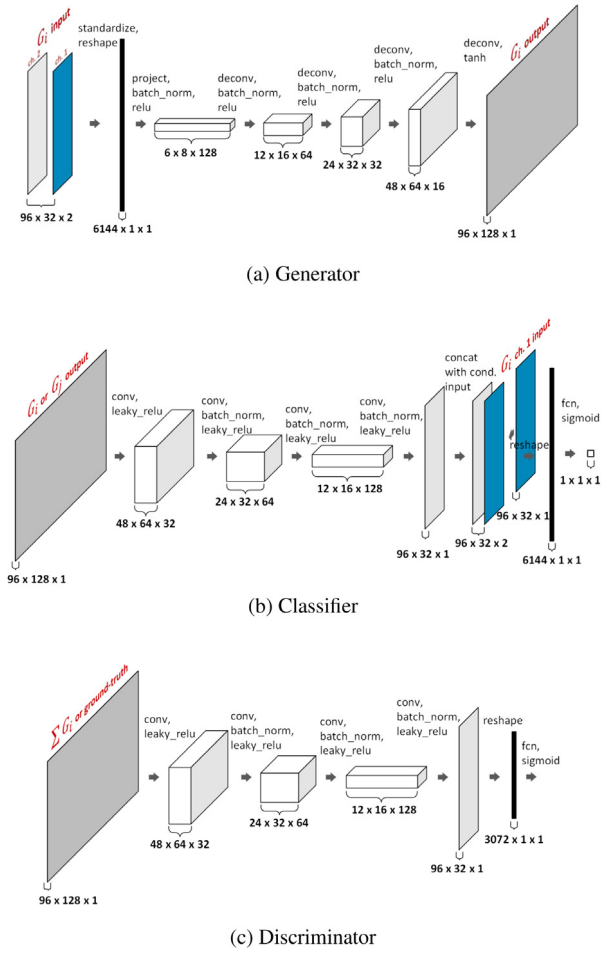


Fig. 6. Detailed architecture of the experimental CMGGAN model.

We calculated the FID distance after each training epoch of the CMGGANs, InfoGANs and CGANs. Lower FID is better, in the sense of better image quality and diversity, and as can be observed in Fig. 8, CMGGAN delivers lower FID for almost all training epochs when compared to both, InfoGAN and CGAN networks.

After each training epoch, CMGGAN network is used to generate images using the radar data from the labeled validation data-set. We applied the Fully Convolutional Network for Semantic Segmentation (FCN) (Long et al., 2015) and the Pyramid Scene Parsing Network (PSPNET) (Zhao et al., 2017) to these generated images, and to the ground-truth images from the data-set to obtain the free drivable space semantic labels. Both, FCN and PSPNET networks were pre-trained on ADE20k data-set (Zhou et al., 2017). These calculated labels and the manually created labels were used to determine the mean Intersection

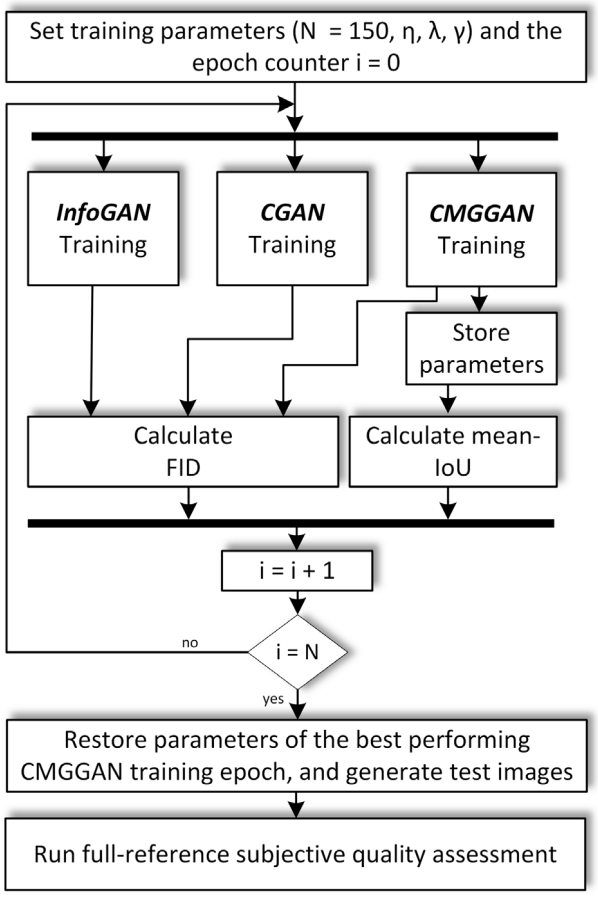
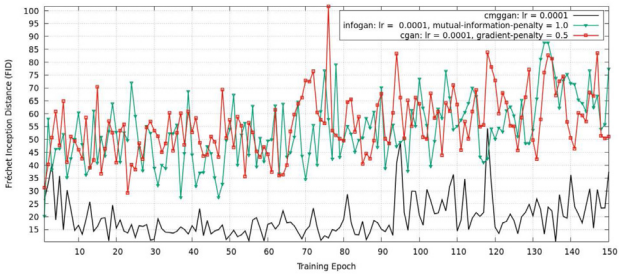
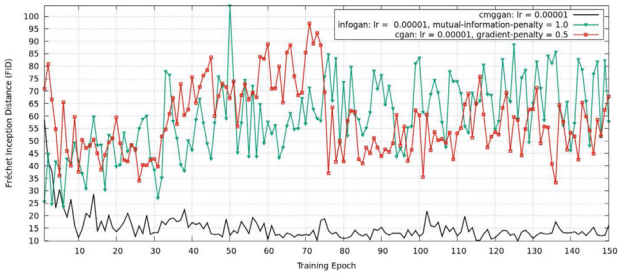


Fig. 7. Model performance assessment procedure.

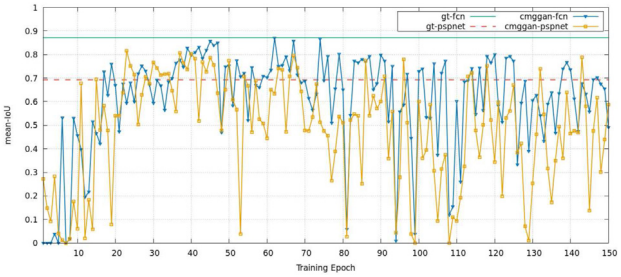


(a) Learning rate $\eta = 0.0001$

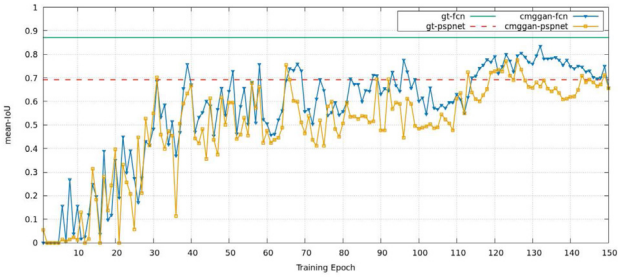


(b) Learning rate $\eta = 0.00001$

Fig. 8. Fréchet Inception Distance (FID) for the CMGGANs, InfoGANs and CGANs, calculated after each training epoch for two best performing training hyper-parameter setups (lower FID is better).



(a) Learning rate $\eta = 0.0001$



(b) Learning rate $\eta = 0.00001$

Fig. 9. Mean-IoU of the PSPNET and FCN semantic labels obtained from the CMGGAN generated image samples, and from the ground-truth images in the validation set.

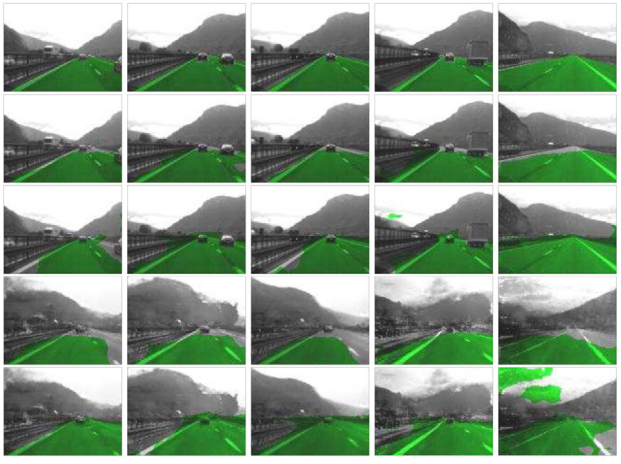


Fig. 10. Manually labeled drivable free space areas and the labels obtained using the semantic segmentation algorithms. From top to bottom, rows are representing: manually labeled ground-truth image, labeled ground-truth images using the FCN, labeled ground-truth images using the PSPNET, labeled generated images using the FCN, and labeled generated images using the PSPNET. Column-wise, generated images are generated from the radar data, which belongs to the same validation set pair as the ground-truth image.

over Union (mean-IoU) score. The validation algorithm first calculates Intersection over Union ($IoU = TP/(TP + FP + FN)$) for each image independently, and then averages the results over the entire validation data-set. The results are presented in Fig. 9. Green and dashed red lines show the mean-IoU for the labels calculated for the ground-truth images by the FCN and PSPNET networks, respectively. These values are constant since the ground-truth images do not change over the training cycles. Although the highest mean-IoU score is obtained after 74 training epochs with the learning rate $\eta = 0.0001$, more stable mean-IoU was obtained with the smaller learning rate $\eta = 0.00001$ and the network trained using this parametrization at epoch 132 is used to generate the image samples used for semantic segmentation. Fig. 10 shows the semantic segmentation results obtained with the FCN

Table 1
General question responses.

	General question score				
	1	2	3	4	5
Count	96	157	177	86	84

Table 2
Vehicle position and basic road features responses.

Answers	Detailed questions			
	Vehicle position	Road curvature	Road width	Dynamic objects
Correct [count]	529	544	516	217
Incorrect [count]	71	56	84	383
Correct [%]	88.2	90.6	86.0	36.2

and PSPNET networks, while Fig. 11 shows the generated images with randomly selected radar grid layers and the corresponding ground-truth images for the two experiments.

As it can be observed, in most cases, the model correctly generates the road width and system-vehicle position. Compared to the ground-truth images that are affected by the high illumination of the camera, the model generates more usable images, as the radar sensor is unaffected by this phenomenon. The overhead signs are not visible on the generated images, as these objects are not reliably detected by the radar due to the low radar mounting position and the sensor's low angular resolution in elevation. As the generators do not have this information, they do not generate such objects in the image.

For the further analysis, we used the CMGGAN generators from the best performing training epoch of the experiment parametrized with the learning rate $\eta = 0.00001$ to generate the image samples with the radar data from the entire test data-set and proceeded with the full-reference subjective quality assessment of these images. A conducted experiment should answer the question whether the generated image contains all necessary road features required to make a decision about the occupancy state of the system-vehicle environment. We created a questionnaire with one general and ten detailed questions for each ground-truth and generated image pair. The general question was: "How well does the generated image (when compared to the given ground-truth scene) represent all road elements?" This question was to be answered with scores between 1 to 5: 1 — there are unacceptable differences, 2 — there are important differences, 3 — there are acceptable differences, 4 — there are slight differences, 5 — there are no noticeable differences. The following detailed questions were to be answered with "yes" and "no":

1. Does the generated image show the correct lateral position of the vehicle?
2. Does the generated image show correct road curvature?
3. Does the generated image show correct road width (all driving and stopping lanes together)?
4. Is the system-vehicle on the generated image in the tunnel?
5. Is the system-vehicle on the ground-truth image in the tunnel?
6. Does the generated image show any overhead obstacles (bridge, traffic sign, and others) in front of the system-vehicle?
7. Does the ground-truth image show any overhead obstacles (bri dge, traffic sign, and others) in front of the system-vehicle?
8. Does the generated image show any other stationary obstacles in front of the system-vehicle?
9. Does the ground-truth image show any other stationary obstacles in front of the system-vehicle?
10. Does the generated image show any dynamic obstacles (other vehicles) in front of the system-vehicle?

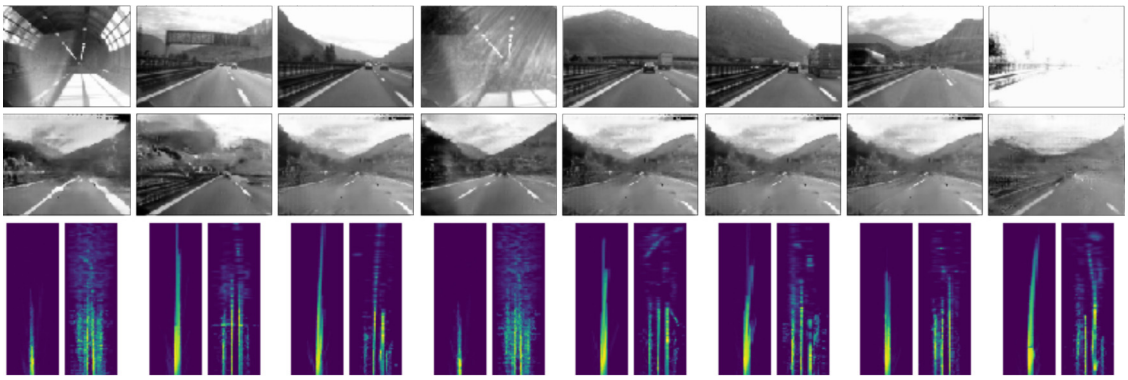
Twelve volunteers participated in the experiment. The participants were instructed which elements of the scene to assess and which to ignore. Each questionnaire consisted of 50 ground-truth/generated image pairs selected randomly from the test data-set. Each pair was accompanied by the above given 11 questions. The time to complete the questionnaire was not limited, and it was required to answer all the questions in the questionnaire. General question responses are summarized in Table 1. It should be noted, that the human subjects rated only 16% of the generated images as having unacceptable differences, when compared to the ground-truth images.

Further on, we analyzed the detailed question responses using two different methodologies. For the unambiguous questions about the system-vehicle position on the road, about the road curvature and the road width, and about the dynamic object appearances we summarized the answers in Table 2 using a simple counting principle. As can be seen, the model delivers excellent results in generating appropriate road features. A lower percentage of correctly generated dynamic objects was expected and desired, as the conditional input variables for the CMGGAN were supposed to contain only detections of the stationary objects. We explain some appearances of the correctly generated dynamic objects due to radar misclassification of the target as a stationary object, or from the image pairs that do not contain any dynamic objects (which caused the assessor to answer "yes" to question 10).

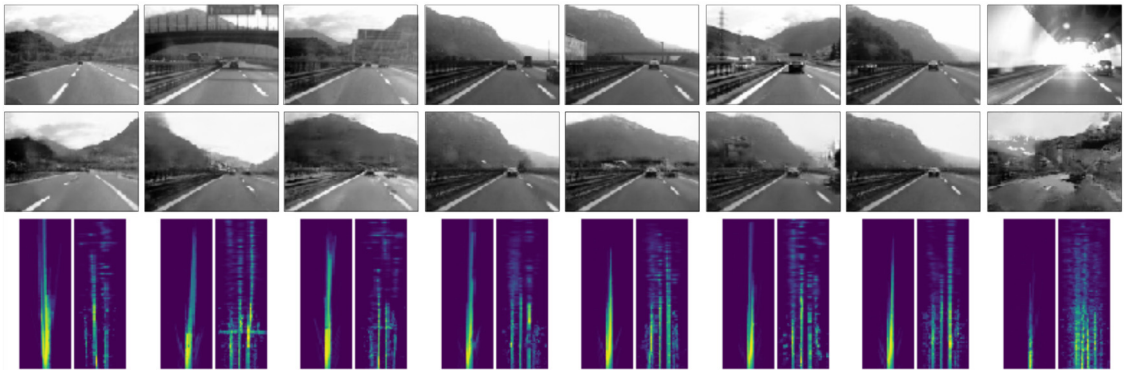
For the rest of the detailed questions responses, we used statistical measures specificity and sensitivity to assess the model performance. The specificity measures the proportion of the negatives that are correctly identified as such, $TN/(TN + FP)$. The sensitivity is equal to the proportion of the positives that are correctly identified as such, $TP/(TP+FN)$. These measures are of crucial importance to the driver assistance systems. Any false positive (Type 1 error) might cause a system function to engage in unexpected vehicle braking, while any false negative detection (Type 2 error) might cause a system function to not respond to a true obstacle. Needless to say, either of the two error types could lead to severe accidents. The obtained results are summarized in Table 3. The proposed algorithm achieves high specificity measures. For the overhead obstacles that do not present a driving obstacle, a low sensitivity is expected, since the conditional input variables (radar data) do not contain elevation measurements. For other stationary objects, there was a higher false positive count than in the other cases, lowering the specificity to an undesirable 0.7940. We assume this problem is partially caused by the radar's ghost detections, caused by the multi-path propagation of a transmitted radar wave or due to the interference from the other radar sensors. On the other hand, this measure can be used to assess the quality of the sensor itself.

5. Concluding remarks

We proposed a novel method for the fusion of automotive radar measurements and camera images. The method is based on the Conditional Multi-Generator GAN unsupervised machine learning algorithm. The key advantage of CMGGAN is its ability to learn to generate different image features when conditioned on a set of meaningful latent variables that target structured semantic information of the data distribution. When conditioned on radar data, CMGGANs generate camera-like images that contain all the environment features detected by the radar sensor. To establish confidence in the soundness and usefulness of the proposed CMGGAN model, we applied the proposed fusion method to a reliable free drivable space detection problem in a highway driving scenario and evaluated it based on the strict experimental setup. When compared to the InfoGANs and CGANs, proposed CMGGANs deliver lower Fréchet Inception Distance for almost all training epochs. Semantic segmentation of generated and camera images was performed using Fully Convolutional Network for Semantic Segmentation and the Pyramid Scene Parsing Network, both pre-trained on ADE20k data-set. Mean Intersection-over-Union was calculated after each training epoch using these calculated semantic labels and manually labeled images from the validation data-set. Obtained mean-IoU scores establish confidence that the images generated from



(a) Learning rate $\eta = 0.0001$, training epoch $n = 74$



(b) Learning rate $\eta = 0.00001$, training epoch $n = 132$

Fig. 11. Generated images from the two experiments with the highest mean-IoU. From top to bottom, on each of the sub-figures, rows are representing: ground-truth images, generated images, and the generator inputs. Column-wise, input radar data belongs to the same test set pair as the ground-truth image.

Table 3
Statistical measures derived from the responses.

	Model statistical measures					
	TP [count]	FP [count]	TN [count]	FN [count]	$\frac{TN}{TN+FP}$ [specificity]	$\frac{TP}{TP+FN}$ [sensitivity]
Tunnel detection	0	5	595	0	0.9917	0.0000
Overhead obstacle detection	5	0	455	140	1.0000	0.0345
Other stationary obstacle detection	0	123	474	3	0.7940	0.0000

the radar data by the CMGGANs can be fused at a pixel level with original camera images of the same scene to increase the robustness of a camera, or their semantic labels can be used to enable an efficient end-to-end semantic segmentation of the radar data, completely avoiding costly and challenging labeling of the radar sensor data.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no conflict of interest.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. arXiv preprint arXiv: 1701.07875.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder–decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2172–2180.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Cho, H., Seo, Y.W., Kumar, B.V., Rajkumar, R.R., 2014. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE, pp. 1836–1843.

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

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009 CVPR 2009 IEEE Conference on. IEEE, pp. 248–255.

Denton, E.L., Chintala, S., Fergus, R., et al., 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems. pp. 1486–1494.

Dubé, R., Hahn, M., Schutz, M., Dickmann, J., Gingras, D., 2014. Detection of parked vehicles from a radar based occupancy grid. In: Intelligent Vehicles Symposium Proceedings, 2014 IEEE. IEEE, pp. 1415–1420.

Goodfellow, I.J., 2017. NIPS 2016 tutorial: Generative adversarial networks. CoRR abs/1701.00160.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637.

Hoang, Q., Nguyen, T.D., Le, T., Phung, D.Q., 2017. Multi-generator generative adversarial nets. CoRR abs/1708.02556.

Lombacher, J., Hahn, M., Dickmann, J., Wöhler, C., 2016. Potential of radar for static object classification using deep learning methods. In: Microwaves for Intelligent Mobility (ICMIM), 2016 IEEE MTT-S International Conference on. IEEE, pp. 1–4.

Long, J., Shelhamer, E., Darrell, T., 2015, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR abs/1411.1784.

Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes, in: ICCV, pp. 5000–5009.

Pagac, D., Nebot, E.M., Durrant-Whyte, H., 1996. An evidential approach to probabilistic map-building. In: Robotics and Automation, 1996 Proceedings. 1996 IEEE International Conference on. IEEE, pp. 745–750.

Pan, J., Canton, C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.a., 2017. Salgan: Visual saliency prediction with generative adversarial networks. arXiv.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. 77, 157–173.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems. pp. 2234–2242.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Wu, H., Siegel, M., Stiefelhagen, R., Yang, J., 2002. Sensor fusion using dempster-shafer theory [for context-aware hci]. In: Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE. IEEE, pp. 7–12.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.