Forecasting Future Instance Segmentation with Learned Optical Flow and Warping

Andrea Ciamarra, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo

University of Florence, Italy
fname.lastnameg@unifi.it

Abstract. For an autonomous vehicle it is essential to observe the ongoing dynamics of a scene and consequently predict imminent future scenarios to ensure safety to itself and others. This can be done using di erent sensors and modalities. In this paper we investigate the usage of optical ow for predicting future semantic segmentations. To do so we propose a model that forecasts ow elds autoregressively. Such predictions are then used to guide the inference of a learned warping function that moves instance segmentations on to future frames. Results on the Cityscapes dataset demonstrate the e ectiveness of optical- ow methods.

Keywords: video prediction · instance segmentation · deep learning

1 Introduction

Understanding urban environments from an ego-vehicle perspective is crucial for safe navigation. This problem can be declined under several points of view, such as detecting entities [9][25], understanding the road layout [2] or predicting the future [21][19][10]. At the same time, di erent modalities can be exploited to understand the environment ranging from RGB, which can produce useful representations such as semantic segmentations [18], to depth or even data produced by more complex sensors such as LiDARs [1], event cameras [25] or thermal cameras [13]. In this paper we focus on predicting future instance semantic segmentations of moving objects, since we believe it is the most informative representation for a machine planning motion. Most approaches are multimodal [8], integrating multiple modalities to achieve the highest accuracy. In general semantic segmentation [20], optical ow [30] and deep features [19] are considered as a source for future prediction. In this paper we study how can we rely on optical ow as unique source of information to forecast instance segmentations. We break down the problem in two steps: optical ow forecasting and learned instance warping. Our contribution is twofold:

- We rst design an optical ow predictor to generate future ows based on convolutional LSTM and a UNet architecture.
- We train a warping neural network, with a specialized loss, that propagates current instance segmentations onto future frames, guided by predicted optical ows.

We show that, by choosing the appropriate loss and by learning a warp operator over noisy autoregressive inputs, we can obtain state-of-the art results on par with more complex multi-modal multi-scale approaches.

2 Related Work

Anticipating the future is a desirable asset for decision-making systems in variegate applications, like autonomous driving and robot navigation. In these scenarios, an arti cial agent should have an intelligent component that forecasts motion of surrounding agents so to navigate safely. Huge progress has been done on anticipating the future by addressing di erent tasks, also exploiting multiple modalities. Early work addressed the video prediction task, where RGB past frames are used to synthesize future frames [24]. A large variety of methods was introduced, including autoregressive models [12], adversarial training [22], bidirectional ow [15], 3D convolutions [32] and recurrent networks [31] to improve prediction accuracy. Instead of predicting raw RGB values, in this work we are interested in directly forecasting more meaningful semantic-level contexts for future frames. Several works have been recently proposed addressing semantic segmentation and instance segmentation forecasting.

Future Semantic Segmentation Scene understanding through future semantic segmentation can help an autonomous agent to take more intelligent decisions, e.g. forecasting speci c category classes like pedestrians [3] or vehicles [4]. Luc et al. [20] proposed to learn semantic-level scene dynamics by mapping semantic segmentations, which are extracted from past frames, to the future through a deep CNN. In this work the authors also shown that directly predicting semantic segmentation is more e ective than predicting RGB frames and then fed them into a semantic segmentation model. Terwilliger et al. [30] designed a ow-guided model, which rst aggregates past ow features using a recurrent network to predict the future optical ow; then, the predicted ow is used to learn a warp layer to map the last semantic segmentation to the future next one. Chiu et al. [5] proposed a teacher-student learning network, which encodes past RGB frames to directly generate future segmentations. Other recent approaches extract features at di erent resolutions from past frames in order to map to the future ones, taking inspiration from the F2F architecture designed in [19]. Saric et al. [27] used deformable convolutions to F2F network at a single-level in order to encode varied motion pattern. In [28], the same authors enriched the features extracted from past images with correlation coe cients between neighbouring frames, and performed the forecasting by warping observed features according to regressed feature ow.

We propose to learn motion dynamics through an optical ow predictive model, using an encoder-decoder model with skip connections followed by a ConvLSTM [34]. Di erently from [30] and [28], we generate future semantic segmentations of moving objects, by aggregating together warped object masks, using the predicted ow and the semantic segmentation as inputs. Speci cally,

our model forecasts single instance segmentations, which allows a ner grained understanding of the scene for tasks such as path planning. In addition, we exploit a fully learnable warping module instead of relying on a di erentiable, yet not-learnable, warping operator as in [30].

Future Instance Segmentation The goal of the future instance segmentation task is to detect and segment individual objects of interest appearing in a video sequence. Semantic segmentation does not account for single objects, since they are fused together by assigning the same category label. Moreover instances can vary in number between frames and are not consistent across a video sequence. Most recent methods generate future instance and semantic segmentations, by forecasting features for unobserved frames given the past ones. For instance, Luc et al. [19] designed a multi-scale approach, named F2F, based on a convolutional network with Feature Pyramid Networks (FPN)[16]. Then, future instance segmentations are detected by running Mask R-CNN[9] from the predicted features, while future semantic segmentations, are computed by converting instance segmentation predictions according to the con dence score, thus proving better qualitative results than [20]. Sun et al. [29], following the same pipeline in [19], employs a set of connected ConvLSTMs [34], so to capture rich spatio-temporal contexts across di erent pyramid levels. Hu et al. [10] proposed a multi-stage optimization framework that exploits auto-path connections so to adaptively and selectively aggregate contexts from previous steps.

Lin et al. [18] designed a three-stage framework, able to forecast segmentation features from a decoder through predictive features outputted by an encoder, instead of directly predict pyramid segmentation features from input frames. Different than F2F [19], future predictions are nally generated by jointly training the whole system using Mask R-CNN [9] and Semantic FPN [14], for semantic segmentations and instance segmentations respectively. Graber et al. [8], proposed a more complete framework to forecast the near future, by decomposing a dynamic scene into *things* and *stu* , i.e. individual objects and background, with multiple training stages and also considering odometry anticipation due to camera motion. In summary, recent works mainly address the future prediction task, both in terms of instances and semantics, following the typical pipeline based on F2F, using recurrent networks and complex connections. The predicted features are then fed in input to di erent segmentation models, so to provide instances and semantic predictions. Although taking several training stages, such frameworks are able to get better results. Our method, instead, proposes to combine the predicted dense optical ow with an instance warping network, without taking several optimization stages. We also prove that the warping model can learn to forecast object motion, from the predicted scene ow and the current semantic map.

3 Our Approach

We introduce a novel framework for future instance segmentation that learns to predict future ow motion patterns. Such predictions are used along with the last

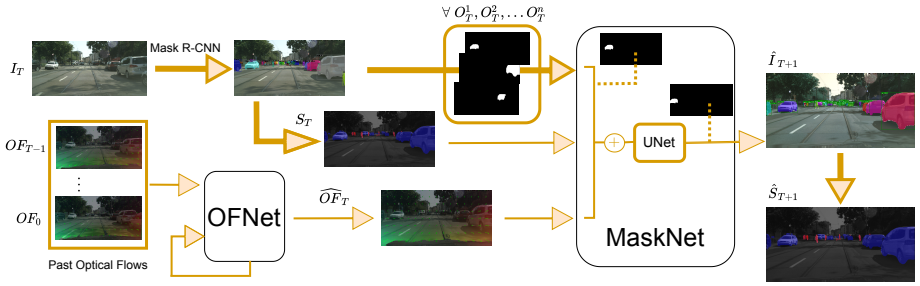

Fig. 1: OFNet forecasts future flows from past ones autoregressively. MaskNet warps instance segmentations to the future with the predicted flow. Future semantic segmentations are obtained by aggregating instances together.

observed semantic segmentation to warp detected objects to future unobserved frames (Fig. 1). This architecture consists of two main components: (i) OFNet, an optical flow predictive model, which forecasts motion dynamics through past optical flows, and (ii) MaskNet, a novel mask prediction model, trained to warp segmentation instances of moving objects to future frames.

Warping is performed by leveraging a region-based loss function instead of the typical cross-entropy loss, so to deal with small parts of interest since the background occupies most of the scene. To identify instances in a video sequence, we run Mask R-CNN [9] for each frame. Then, we exploit a bounding box association-based tracker based on [7] to generate object tracks and link instances for the sequence lifespan. Inspired by [19], the proposed approach is also able to generate semantic segmentations by fusing instance mask predictions according to a rescoring function balancing detector score and mask size.

3.1 Optical Flow Forecasting

The first key part of the framework is OFNet, the optical flow prediction model. It is a supervised network that learns to generate future flows based on past ones. OFNet takes T dense optical flows from consecutive past frames and predicts the flow for the following time step.

Each optical flow is first estimated from consecutive pairs of frames, by running FlowNet2 [11], that produces a dense 2D flow field, with $(u; v)$ horizontal and vertical displacements for each pixel. Then, we learn motion features from each of these optical flows through a time distributed UNet [26], a fully convolutional encoder-decoder network with a contracting path and an expanding path, connected with skip connections. From the original UNet we consider the entire structure up to the final prediction layer, so to output features with 64 channels (Fig. 2(a)). The sequence of extracted features generated by UNet are fed to a ConvLSTM [34] with 3 3 kernel followed by a 1 1 convolution, so to reduce the output channels to 2 as in a flow field. We train OFNet to forecast an output sequence shifted by one step ahead with respect to input, i.e. by generating

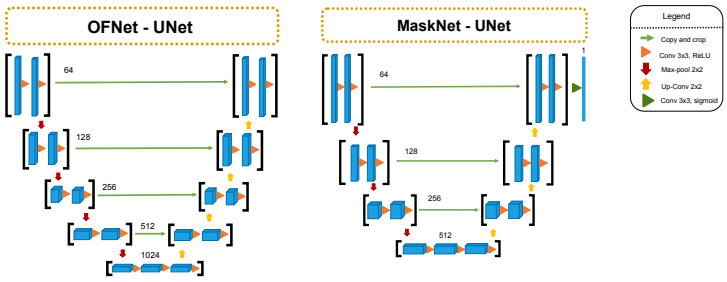

Fig. 2: UNet architectures used as feature extractor (a) to generate 64-channel ow features in OFNet, and (b) to produce future binary masks in MaskNet.

the next ow for each corresponding input. This provides supervision for each timestep and allows us to use it in an autoregressive fashion.

Formally, let OF_i be the optical ow produced by FlowNet2 given two consecutive frames X_i and X_{i+1}. We denote with F_i the motion features extracted from OF_i using our UNet-based backbone. We extend such notation referring to $F_{t:T}$ as the sequence of motion features spanning from t to T. Likewise, we use the same notation for optical ow sequences $OF_{t:T}$. At training time, given a sequence of T ow features $F_{t:t+T}$, OFNet learns to generate a sequence of T consecutive optical ows shifted by one timestep $\widehat{OF}_{t+1:t+T+1}$. Instead, at inference time we only retain the last prediction of the output sequence $\widehat{OF}_{t+T+1}$, that encodes the scene motion towards the rst unseen future frame. OFNet is trained using an L2 loss function on the output sequence L_{ow}:

$$L_{\text{ow}} = \frac{1}{T} \sum_{k=1}^{T} \frac{\left\| OF_k - \widehat{OF}_k \right\|^2}{H \; W \; 2} \tag{1}$$

where T is the sequence length and H and W are the height and width of the feature map. We set $T = 6$ for all our experiments to provide su cient information about past dynamics.

3.2 Instance Mask Forecasting

After generating future optical ows, we forecast future instance segmentations. To do that, we designed MaskNet, a novel model which leverages motions encoded in the predicted ows and the semantic segmentation of an observed frame, which provides context about the objects in the scene. Instead of globally warping semantic segmentations as done in [30], MaskNet is trained to understand future positions of individual objects. We aggregate information from the present frame X_t, i.e. the semantic segmentation of the scene S_t and a binary mask of

the instance to be warped M_t^j, and from the estimated future, i.e. the predicted flow field OF_t. The data is concatenated channel-wise and fed to the network, which generates the future instance mask $\hat{M}_{t+1}^j$ located in frame X_{t+1}.

As in OFNet, we use a UNet backbone for MaskNet but we reduce its encoder-decoder structure by one level to avoid overfitting (Fig. 2(b)). After aggregating $\{OF_t, S_t, M_t^j\}$, MaskNet takes as input a tensor of size $H \times W \times 4$ and produces as output a binary map $H \times W \times 1$, where 0 corresponds to background and 1 to the forecasted instance mask. To binarize the output we use a sigmoid activation and we threshold at 0.5. Since an object mostly occupies a very small region compared to the input resolution, we are in the typical case of class unbalance. To address this issue, rather than training MaskNet with a cross-entropy loss, we use a Dice loss [23]. Dice loss is a region-based loss that helps to better focus on overlapping regions in two images, which in our case are the predicted binary mask and the corresponding groundtruth instance. Thus, considering an N pixel image, we employ the mask loss L_{mask} defined through the Dice coefficient D:

$$L_{\text{mask}} = 1 - D = 1 - \frac{2 \sum_{i=1}^{N} \hat{M}_i M^{GT}{}_i}{\sum_{i=1}^{N} \hat{M}_i^2 + \sum_{i=1}^{N} M^{GT}{}_i^2} \tag{2}$$

where the Dice coefficient $D \in [0,1]$ measures the similarity between the predicted mask $\hat{M}$ and the groundtruth one M^{GT}. Note that to compute the loss, we use the output of MaskNet after the sigmoid activation and before the final thresholding in order to work with real values in $[0,1]$. Since our model works with single instances, we do not rely on a dedicated network to generate semantic segmentations of the image as most state of the art methods [8][30][18]. Instead, we simply group all object masks together maintaining their category label.

4 Experiments

We conduct experiments on the Cityscapes dataset [6], which contains recordings from a car driving in urban environments for a total of 5000 sequences, divided in 2975 for train, 1525 for testing and 500 for validation. Every sequence is long 1.8s and consists of 30 frames at resolution 1024×2048. Groundtruth semantic and instance segmentations are available for the 20-th frame of each sequence. Our method is tested on 8 moving object categories: *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle*, *bicycle*. We evaluate the future predictions, both instance and semantic segmentations, 3 and 9 frames ahead, respectively denoted as short-term (about 0.17s later) and mid-term (about 0.5s later). Such evaluations are performed on annotated frames in the validation set, as done in literature. Predicted optical flows are obtained, by initially feeding $T = 6$ consecutive past optical flows $OF_{t-6:t-1}$ to OFNet and then, autoregressively forecasting each future flow OF_{t+k} with $k = 0,1,\dots,n$ ($n = 2$ for short-term and $n = 8$ for mid-term). Future semantic segmentation is evaluated by measuring the intersection-over-union (IoU) between predictions and the corresponding ground truth for each class and averaging among classes. Future instance segmentation instead is evaluated through two standard Cityscapes metrics: AP50

Table 1: E ect of di erent training regimes for MaskNet. Warping instances with predicted ows is included as a baseline.

Method	Train Data		Finetuned	Short term $t = 3$			Mid term $t = 9$		
	FlowNet	OFNet	Layers	AP	AP50	IoU	AP	AP50	IoU
Warping	7	7	-	15.9	34.3	60.7	2.8	8.8	39.4
MaskNet	3	7	-	18.7	39.4	65.5	4.0	12.3	44.5
	7	3	all	18.5	39.3	65.2	4.5	14.9	41.9
	3	3	3	18.7	39.2	65.7	5.1	16.0	44.8
	3	3	2	**18.9**	**39.5**	**65.9**	**5.9**	**17.4**	**45.5**

and AP. AP50 is average precision where an instance is correctly counted if it has at least 50% of IoU with the groundtruth. AP is average precision, averaged over ten equally spaced IoU thresholds from 50% to 95%. Since these metrics require a con dence score for each prediction, we use the score provided by the detector for each object. However, small objects are likely to be false positives generated by the detector. To mitigate this issue, we reduce the score by 0.5 for objects smaller than 64x64px and by 0.3 for objects smaller than 128x128px.

4.1 Implementation Details and Ablation Study

We compute the optical ow at every pair of consecutive frames for each video sequence, both for training and validation sets, through FlowNet-c with pretrained weights from FlowNet2 [11]. Frames are resized to 128 256, as in [30]. We detect all objects from each video frame, by running the Mask R-CNN model provided in Detectron2 [33], pretrained on COCO [17] and netuned on Cityscapes. We train OFNet with Adam and learning rate 0.0001. Our architecture is inspired by UNet, but in order to perform regression we change the number of channels to 2 in the last 1 1 convolutional layer and we use a linear activation function. We set the batch size to 3 and we employ the L2 loss as in Eq. 1 and we train the network for 100 epochs. MaskNet instead is trained for 4 epochs using Adam with learning rate 0:0001. Inputs are resized to 256 512. First, we pretrain the model using optical ows computed with FlowNet [11] instead of the predicted ones. Then, we netune for 3 additional epochs the last two layers using the predicted ows autoregressively, for both short-term and mid-term models. This two-step optimization allows the model to learn how to warp instances using clean ows and then to adapt itself to handle noisy ow elds.

We observe that using a Cross-Entropy loss instead of a Dice loss, MaskNet converges to trivial solutions predicting background for every pixel of the frame. This underlines the importance of using a region based loss for this kind of task.

In Tab. 1 we conduct an ablation study by exploring di erent training strategies. As a rst control experiment, we output future segmentations by directly warping masks with ows predicted by OFNet. In this case MaskNet is not used. Then we evaluate the e ect of pretraining MaskNet using ground truth ows generated by FlowNet [11]. Di erent ne-tuning strategies are also analyzed.

Table 2: Flow prediction MSE up to mid-term (t+9) on Cityscapes val set. We also show a break down of MSE for the u and v optical ow components.

	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9
MSE	0.36	0.54	0.60	0.78	0.84	0.99	1.08	1.28	1.19
MSE_u	0.45	0.67	0.70	0.94	1.04	1.24	1.31	1.64	1.55
MSE_v	0.26	0.41	0.50	0.63	0.63	0.75	0.84	0.91	0.82

As expected using a learned warping function, MaskNet, yields superior results with respect to simply warping masks. It also emerges that the two step training is fundamental. Training MaskNet only on precomputed ows yields good results at short term but exhibits a severe drop in performance at mid-term, since the model is not trained to deal with noisy predictions as inputs. Similarly, using only predicted ows, i.e. without pretraining, the model fails to warp masks appropriately up to mid-term. We also study the e ect of using OFNet in an autoregressive way or by feeding FlowNet ows to it in a teacher-forcing way. This allows us to better understand the impact of ow quality on MaskNet's ability to warp masks. Interestingly, when OFNet is fed autoregressively with predicted ows, MaskNet exhibits a large improvement in average precision for mid-term predictions. In fact, using a teacher forcing approach, i.e. feeding only clean ows to OFNet, the produced ows are not su ciently realistic. This is con rmed by the fact that this model performs similarly to the one trained only on real ows. Finally, we compare performances by netuning the last 3 layers instead of 2. We found that 2 layers to train is the best choice.

5 Results

We rst analyse the prediction capabilities of OFNet. Since we want to use it to generate future ows up to mid-term (9 frames in the future), we show in Tab. 2 how error accumulates while predicting ows autoregressively. As expected, the MSE values increase as the time horizon shifts forward, although the error accumulation appears to be bounded by a linear growth. Interestingly, the horizontal component u exhibits a higher error. This is caused by the fact that most objects move horizontally in front of the camera capturing an urban scene from an ego-vehicle perspective.

We then provide in Fig. 3 a qualitative analysis of the warping capabilities of MaskNet feeding clean future ows using FlowNet2 as an oracle. In the following we will refer to this model as *MaskNet-Oracle*.

This model can provide accurate results among di erent categories. Interestingly, it can correctly warp small non-rigid parts of an instance, like legs, as well as objects at di erent scales like approaching cars. These results prove that MaskNet can understand most motions of individual objects in a scene by simply exploiting optical ow, without further motion information. MaskNet-Oracle, being fed with ground truth optical ows, acts as an upper bound of

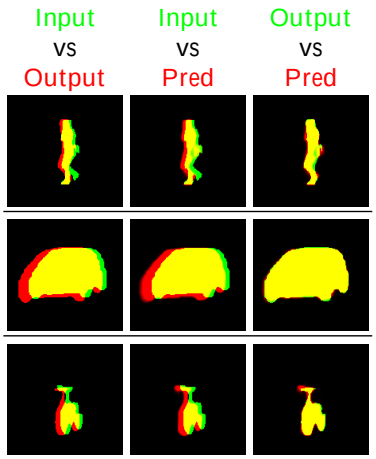

Fig. 3: Instance predictions: input vs gt output, input vs prediction, gt output vs prediction. Mask intersections in yellow.

Method	Short term		$t = 3$	Mid term		$t = 9$
	AP	AP50	IoU	AP	AP50	IoU
Mask RCNN oracle	34.6	57.4	73.8	34.6	57.4	73.8
MaskNet-Oracle	24.8	47.2	69.6	16.5	35.2	61.4
Copy-last segm. [19]	10.1	24.1	45.7	1.8	6.6	29.1
Optical- ow shift [19]	16.0	37.0	56.7	2.9	9.7	36.7
Optical- ow warp [19]	16.5	36.8	58.8	4.1	11.1	41.4
Mask H2F[19]	11.8	25.5	46.2	5.1	14.2	30.5
S2S[20]	-	-	55.4	-	-	42.4
F2F[19]	19.4	39.9	61.2	7.7	19.4	41.2
LSTM M2M[30]	-	-	65.1	-	-	46.3
Deform F2F[27]	-	-	63.8	-	-	49.9
CPConvLSTM[29]	22.1	44.3	-	11.2	25.6	-
F2MF[28]	-	-	67.7	-	-	54.6
PSF[8]	17.8	38.4	60.8	10.0	22.3	52.1
APANet[10]	23.2	46.1	64.9	12.9	29.2	51.4
PFL[18]	24.9	48.7	69.2	14.8	30.5	56.7
MaskNet (Ours)	19.5	40.5	65.9	6.4	18.4	45.5

Table 3: Comparison results for future instance segmentation (AP and AP50) and future semantic segmentation (IoU) of moving objects on Cityscapes val set.

our system. We also evaluate a Mask R-CNN oracle, by directly running Mask R-CNN on future frames. We compare our approach, that uses the predicted optical ows from OFNet, with such upper bounds and with recent approaches from the state of the art (Tab. 3). We exceed some state of the methods, such as F2F and PSF [8] at short-term, while mid-term predictions are very close to the feature-to-feature approach introduced by Luc et al. [19]. This con rms our choice of using motion cues, instead of learning di erent Feature Pyramid Networks. However, recent frameworks, such as CPConvLSTM [29], APANet [10] and PFL [18], reported better results, where more complex information from intra-level and inter-level connections are learned across FPNs from past frames. This requires several training stages to improve both short and mid predictions. We also generate future semantic segmentations for moving objects (Tab. 3). We obtain the future semantic segmentations by grouping by category the predicted instance masks. MaskNet obtains good results and overcomes most of those state-of-the-art approaches that provide forecasts in terms of both instance and semantic segmentations. In general our approach performs well at short-term, but gets worse at mid-term compared to methods that predict future FPNs after several training stages and learning speci c connections, like the auto-path connections as in APANet [10], or that use segmentation models properly trained on predicted features, as done in PFL [18]. Some qualitative results are shown for instance and semantic segmentation predictions, both at short-term and mid-term, in Fig. 4.

Overall, compared to the other methods, our approach is simple and without complex optimization stages can reach competing performances. Nonetheless,

Fig. 4: Visualization results of future predictions on Cityscapes val set.

our approach underlines the e cacy and the limitations of relying solely on optical ow data, which is easy to obtain even with any typical on-board sensor whether it is as RGB camera, an event camera or a LiDAR. However, we believe that there is still margin for improvement using a ow-based predictor like ours. This stems from the results of our MaskNet-Oracle, as using clean ows estimated by FlowNet2 leads to overcome all state-of-the-art results at both short-term and mid-term. We believe that integrating information such as depth-maps could lead to signi cant improvements in the state of the art. However, this is out of the scope of our paper and we leave it as future work.

6 Conclusions

In this work we addressed the future instance segmentation task using predicted optical ows and a learned warping operator, which also allows to produce future semantic segmentations without additional training. We designed a framework made of a predictive optical ow model able to forecast future ows given past ones through a UNet+ConvLSTM architecture. A second component, MaskNet, warps observed objects to future frames, using the current semantic map and the predicted optical ow. We conducted experiments on Cityscapes, demonstrating the e ectiveness of optical ow based methods.

Acknowledgements This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media

References

1. Beltran, J., Guindel, C., Moreno, F.M., Cruzado, D., Garcia, F., De La Escalera, A.: Birdnet: a 3d object detection framework from lidar information. In: 2018 21st