IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE VEHICULAR TECHNOLOGY SOCIETY SECTION

Received October 10, 2021, accepted October 22, 2021, date of publication October 26, 2021, date of current version November 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123169

LiCaNext: Incorporating Sequential Range Residuals for Additional Advancement in Joint Perception and Motion Prediction

YASSER H. KHALIL, (Member, IEEE), AND HUSSEIN T. MOUFTAH, (Life Fellow, IEEE)
School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Hussein T. Mouftah (mouftah@uottawa.ca)

This work was supported by the Canada Research Chairs Fund Program and Natural Sciences and Engineering Research Council of Canada through the Discovery Grant Project under Grant RGPIN/1056-2017.

ABSTRACT Autonomous driving can obtain accurate perception and reliable motion prediction with the support of multi-modal fusion. Recently, there has been growing interest in leveraging features from various onboard sensors to enhance the primary stages of autonomous driving. This paper proposes *LiCaNext* to capture additional accuracy advancements in joint perception and motion prediction while maintaining real-time requirements. *LiCaNext* is the *next* version of LiCaNet, which fuses LIDAR data in both bird's-eye view (BEV) and range view (RV) representations with a camera image. In contrast to LiCaNet, we introduce sequential range residual images into the multi-modal fusion network to further improve performance, with minimal increase in inference time. Employing sequential range residual images has a substantial direct impact on motion prediction and positively influences perception. We provide an extensive evaluation on the public nuScenes dataset. Our experiments show that incorporating sequential range residuals secures significant performance gain, with monotonic progress for a larger number of exploited residuals.

INDEX TERMS Autonomous driving, deep learning, motion prediction, multi-modal fusion, perception, residual image, sensor fusion.

I. INTRODUCTION

Perception and motion prediction components are critical to the safety of autonomous driving and must not be overlooked. Recently, researchers invested an ample amount of time and momentum towards improving the safety of autonomous driving by enhancing the accuracy and reliability of its primary components [1]–[7]. Currently, the main research focus is on exploiting multi-modal fusion to develop superior versions of perception and motion prediction components. The multi-modal fusion technique leverages data extracted from a diverse range of sensors commonly deployed on an autonomous vehicle to 1) better infer the current state of its surroundings and 2) accurately predict the dynamicity of this state in the near future. The utilization of multi-modal fusion in perception and motion prediction proved its competence in enhancing performance [1], [3], [5], [6]. The prominence of multi-modal fusion stems from the fact that the generated features are rich and complete. These features constitute the complementary properties of all fused data representations and alleviate the inherent constraints of individual representations.

This work focuses on developing a joint perception and motion prediction model for autonomous driving by combining data from LIDAR and camera sensors. The two most common representations of LIDAR data are bird's-eye view (BEV) [3], [4] and range view (RV) [2], [5]. We proposed LiCaNet [1] in our earlier work, which tackled the same problem and produced challenging results. LiCaNet formulates its LIDAR data in BEV and RV representations. The BEV input is composed of a historical sequence of LIDAR data, while the RV and the camera input images represent only the current frame. Inspired by [8], we propose *LiCaNext*, the *next* version of LiCaNet, which expands on the multi-modal fusion network by incorporating sequential range residual images.

Range residuals are computed using the frame differencing technique [9], a pixel level comparison between the current and previous frames. Fig. 1 illustrates an example of

The associate editor coordinating the review of this manuscript and approving it for publication was Jie Gao.

sequential range residual images. A residual image captures rich temporal information of objects, which is vital for identifying motion in a scene, leading to foreground-background differentiation. Object motion segmentation is identified as one of the fundamental requirements for real-world applications, including visual surveillance [10], traffic control [11], autonomous driving [3], and much more. An active research area in scene understanding for autonomous driving is moving object segmentation (MOS) [8], [12]–[15]. MOS is a class-agnostic approach that detects and localizes motion in scenes. Furthermore, MOS can detect unseen objects (e.g., rare animals or construction vehicles) since it relies on motion cues rather than semantics. Typically, the ability to detect the dynamicity of the surrounding environment is pivotal for the safety of autonomous driving because it enables the prediction of objects' future states and path-planning. The terms motion detection and motion segmentation are used interchangeably in this work.

A common challenge in motion detection is the rapid variation in illumination, such as the sudden appearance of clouds in the sky. In reality, under sharp changes of lightning conditions, static pixels are expressed by different intensities from the rest of the background pixels; thus, incorrectly classifying them as foreground. Fortunately, our computed range residuals are invariant to changes in illumination because range images only contain distance values to objects, and intensity values are excluded. Consequently, motion detection performance remains unaffected under adverse illumination situations if temporal information is extracted from range residual images. Accordingly, extracting rich temporal information from range images and integrating them into the multi-modal fusion network permits us to obtain an effective and efficient motion detection invariant to changes in intensities.

The RV input image of our previous multi-modal fusion network lacks temporal information. It merely consists of spatial information, including the range, intensity, and height of objects in the surrounding environment. After inserting sequential range residual images in RV form, our RV input images now generate features that embrace spatio-temporal information. Even though utilizing sequential range residuals is a modest adjustment to our multi-modal fusion network, the boost in perception and motion prediction accuracy that results from this simple addition is substantial. Many models exist in the literature that offers additional performance advancement, in these two critical components, compared to their predecessors; nonetheless, the extra computation introduced is relatively high compared to the additional accuracy attained. LiCaNet model, an enhancement to MotionNet [4], achieved an increase of 2.1% in perception and a mean error drop of 23.1mm in motion prediction for the fast category (pixels having speed $> 5m/s$); with a rise of 7.6ms in inference time. On the other hand, our proposed *LiCaNext* obtains an increase in inference time of only 1.6ms while achieving a perception gain of 1.6% and a mean error decrease of 73.7mm in motion prediction for the fast category. Thus, a simple modification applied to a model that leads to a significant progression in accuracy with a limited increase in computational time is considered a significant contribution.

Accordingly, alongside the temporal information fed into the BEV module of our multi-modal fusion network, the range residuals also hold the dynamicity of the surroundings. This redundant temporal information strengthens the model's confidence in differentiating between foreground-background objects, accurately predicting motion, and enhancing perception. Furthermore, the increase in inference time incurred due to the addition of residual images is minimal. Fig. 2 demonstrates the methodology of inserting sequential range residuals into our multi-modal fusion network.

Overall, the features generated by our proposed *LiCaNext* combines the 1) physical object dimensions and temporal information represented in BEV images, 2) occlusion information characterized in RV form, 3) motion cues embodied in both range residuals and RV forms, and 4) rich semantics of the surrounding environment signified in a camera image. Finally, these generated features are fed into MotionNet backbone network [4] to perform accurate pixel-wise joint perception and motion prediction in real-time.

The key contribution of this paper is the incorporation of residual images into the fusion process of BEV, RV, and camera images. The generated high-quality and complementary features are then fed into the backbone network to accomplish state-of-the-art pixel-wise joint perception and motion predictions in real-time. *LiCaNext* has been tested and evaluated on nuScenes dataset [16]. *LiCaNext* achieves superior results to LiCaNet, proving that incorporating residual images enhances joint perception and motion prediction performance. Moreover, the conducted experiments reveal that the higher the number of residual images injected in the multi-modal network, the greater the performance gain. Additionally, results show that the most significant improvement is recorded for small and distant objects. We believe that our proposed *LiCaNext* is the first existing multi-modal network to fuse sequential range residual images with multi-modal features to perform accurate pixel-wise joint perception and motion prediction in real-time.

The rest of the paper is structured as follows. A brief review of related work is discussed in Section II. Our proposed approach is presented in Section III. Section IV describes the experimental results. Lastly, Section V concludes the paper.

II. RELATED WORK

This section provides a brief review of works that target motion object segmentation for autonomous driving. In addition, we present a concise overview of works that address perception and motion prediction. Motion segmentation is defined as the process of detecting motion at a pixel level, while motion prediction is forecasting the future motion of objects. In our work, we perform pixel-wise motion prediction.

A. MOVING OBJECT SEGMENTATION

Early motion detection architectures relied on a geometric understanding of the scene [17]. Such architectures face difficulty adapting to challenging situations as their designs are complex enough and are created exclusively for specific challenging scenarios. Recent advancements in learning-based approaches lead to massive progress in motion detection [8], [12], [13], [18]–[22]. SMSnet [18] is a method that leverages a convolutional neural network (CNN) and depends on two sequential camera images to perform pixel-wise category labeling and motion detection. MODNet [19] is another CNN model that fuses motion and appearance cues to perform joint detection and motion segmentation. One camera sensor is used in [18], [19] to perform motion segmentation on vehicles only. Unfortunately, it is unsafe to depend merely on a camera sensor for performing MOS because the semantics of the image degrades sharply under low-quality illumination conditions. Conversely, Dewan *et al.* [20] proposed a LIDAR-based method that depends on two sequential scans to perform a pointwise segmentation classifier distinguishing foreground objects from static background. The model in [20] makes use of up-convolutional networks to accomplish the desired task. Moreover, Chen *et al.* [8] proposed real-time class-agnostic motion segmentation using sequential LIDAR scans. The input to the CNN network in [8] is a combination of RV image, representing the current LIDAR scan, and range residuals computed from historical LIDAR frames.

Recent works are targeting the fusion of multi-modal data to attain more robust motion segmentation. RST-MODNet [12] is a CNN architecture that leverages the fusion of sequential camera and optical flow images to achieve real-time motion detection. FuseMODNet [13] is another real-time CNN architecture; however, it depends on combining LIDAR and camera images to capture motion information. Additionally, Mohamed *et al.* [21] developed a real-time CNN architecture, for instance-level class-agnostic motion segmentation with a camera and optical flow images as input. Unlike the earlier versions, [21] improves the diversity of moving objects by adding four additional classes instead of just vehicles. Lastly, BEV-MODNet [22] is another enhancement that investigated the idea of learning motion detections directly on the BEV space. In [22], a deep network is designed with a two-stream RGB and an optical flow fusion architecture.

B. PERCEPTION AND MOTION PREDICTION

The development of perception and motion prediction for autonomous driving has picked a staggering pace in the past few years. Plethora of methodologies has been explored to enhance perception and motion prediction performance. The majority of works available in the literature used single input representation to address this task [4], [7], [23]–[31]. Recently, applying multi-modal feature fusion has sparked a lot of interest from the research community in autonomous vehicles.

To begin with, LiRaNet [6] model fuses instantaneous velocity information of RADAR along with LIDAR data and high-definition (HD) maps to perform perception and prediction. Fadadu *et al.* [5] define a unified architecture that incorporates two views of LIDAR data (BEV and RV), camera, and HD maps for advanced object detection and trajectory prediction. Another fusion method attempts to integrate the outcomes of MotionNet with BEV images to perform efficient and safe autonomous driving in an urban environment by training a reinforcement learning model [3]. Khalil *et al.* [2] put forward a multi-view LIDAR-based fusion network to enhance pixel-wise joint perception and motion prediction compared to MotionNet baseline. The two input LIDAR representations employed in [2] are BEV and RV images. Lastly, recently proposed LiCaNet [1] extends [2] with camera image fusion. LiCaNet records excellent performance for both perception and motion prediction compared to its predecessor.

In comparison to the multi-modal fusion networks mentioned above, we propose *LiCaNext*, a buildup on LiCaNet model where the multi-modal fusion network is expanded to involve range residuals in RV representation. In addition to the temporal information provided by the historical sequence of BEV images, employing motion cues encoded in sequential range residuals reinforces the richness and completeness of the generated features. Therefore, we feed our multi-modal fusion network redundant temporal information in RV form, allowing us to exploit spatio-temporal information in both BEV and RV representations.

III. PROPOSED METHODOLOGY

A. INPUT REPRESENTATION

All input representations in *LiCaNext* are the same as LiCaNet, except for the newly incorporated range residuals.

1) BIRD'S-EYE VIEW

The nuScenes dataset consists of several scenes. We break down each scene into clips, and each clip consists of a current frame and 4 prior frames. The current frame in each clip is sampled at 2Hz for training and 1Hz for testing. The time span between the sampled frames in a clip is 0.2s. Therefore, our BEV input representation constitutes a current frame and 4 previous frames synchronized to the current frame. In terms of height, a range of 5 meters in the z-axis of each frame is encoded in 13 channels. Moreover, each channel has length and width covering a range of 64m in each direction and is encoded in a 2D image of dimensions 256×256. More specifically, the specified range in the xyz-direction is $[-32, 32] \times [-32, 32] \times [-3, 2]$ m, respectively. The resolution of each cell in the BEV image is defined as $(\Delta x, \Delta y, \Delta z) = (0.25, 0.25, 0.4)$ m, where x, y, z denotes the xyz-axis. Therefore, the BEV images in *LiCaNext* are of dimensions $256 \times 256 \times 13 \times 5$.

2) RANGE VIEW

Unlike BEV, no historical information is used to formulate the RV input image. The current frame in RV form is encoded

in 4 channels: range r, height, intensity, and a binary flag indicating whether the cell has a valid value. An empty cell is symbolized in the flag channel by a value of -1 and 1 otherwise. Indeed, a value of -1 is reflected in all 4 channels for an invalid cell. The width and height of the RV image are set to be 1024×32, respectively. Generally, the height of the RV image matches the number of laser beams emitted by the LIDAR sensor. The LIDAR sensor embarked in nuScenes dataset has 32 laser beams.

3) RANGE RESIDUALS

As aforementioned, the main idea of this work is to append sequential range residuals into the multi-modal fusion network, to push further the performance of the joint perception and motion prediction model. The range residuals in RV form are computed by subtracting range images of both current and previous frames. Following [8], the first step towards computing a range residual image is to extract the range images of the two frames separately. The second step is to compute the residual values by subtracting the two range images from each other at a pixel level. Only valid pixels in both range images are considered for computing the residual values of the corresponding pixels. The residual values of all other pixels are set to 0.

However, before computing the range images of the two frames, we must synchronize the previous frame to the current frame. The synchronization process is done by transforming the point cloud of the previous frame into the coordinate system of the current frame. This stage is necessary to counterfeit the ego-motion (motion of the autonomous vehicle). The final step is to normalize the residual images using Eq. (1).

$$d^{\,0}_{n,m} = \frac{\left|\Delta r^{0}_{n,m}\right|}{r^{0}_{m}} \quad \text{and} \quad \Delta r^{0}_{n,m} = r^{0}_{m} - r^{n}_{m}, \tag{1}$$

where $\Delta r^{0}_{n,m}$ is the non-normalized residual image between the current range image (r^0) and the previous n^{th} transformed range image (r^n). Whereas, r^{0}_{m} and r^{n}_{m} represent the range value at cell m of the current and the n^{th} frames, respectively. Lastly, $d^{\,0}_{n,m}$ signifies the residual value at cell m between the 0^{th} and the n^{th} frames. The dimension of each residual image is the same as the individual channels of the RV image (1024×32). If multiple residual images are to be fused in the multi-modal fusion network, then they are stacked on top of each other. For example, if 4 residuals are fused, then the input representation of the sequential range residual images becomes $1024 \times 32 \times 4$.

Fig. 1 provides examples of normalized sequential range residual images at different timestamps. The current and the previous frames are sampled using the same configuration adopted for formulating the BEV input. The range image of the current frame and its corresponding labels, both in RV form, are included in Fig. 1 for reference purposes. The labels are color-coded to easily distinguish and locate the different objects in the range and residual images. The color black is assigned for background, blue for vehicles, red for pedestrians, green for bikes, and brown for all other objects available in the surrounding. The physical appearance of some objects in the label image does not reflect their actual shape due to the scale variance issue in RV images. For instance, some pedestrians are rendered by few dots, whereas, looking at the pedestrian located in the middle of the label image, it is evident from its appearance that it is a pedestrian. Typically, only a few LIDAR points represent distant objects. Thus, when transformed into RV form, their appearance is limited to a couple of pixels, so the appearance will not reflect the object's actual shape. The location of objects relative to the LIDAR sensor is visible in Fig. 2 as it comprises the labels of the current frame in BEV form. Furthermore, Fig. 2 also includes an RGB image displaying the front-side of the current scene within the camera field-of-view (FOV).

Interpreting the residual images in Fig. 1, one can undoubtedly observe the dynamicity of some objects due to their high motion. In comparison, other objects have weak displacement representation because of their moderate motion. In contrast, the remaining objects have void motion as they are static. In particular, a vehicle could be parked on the side of the road, waiting for a traffic light, or even pedestrians could be standing still waiting to cross the road.

4) CAMERA

The front camera in nuScenes dataset captures RGB images with dimensions $1600 \times 900 \times 3$.

B. LiCaNext ARCHITECTURE

The *LiCaNext* architecture consists of four modules: BEV, RV, residual, and camera. The architecture scheme is presented in Fig. 2. The flag channel of the RV image is not depicted in the figure. The addition of the residual module into the LiCaNet architecture leads to *LiCaNext*. A total number of 5 sequential frames are used to represent *LiCaNext* input. The current frame is used to construct: one BEV image (13 channels), one RV image (4 channels), and a single RGB camera image (3 channels). Furthermore, each of the 4 previous frames generates one BEV image (13 channels) and one residual image (1 channel). Therefore, 5 BEV images are stacked to form the BEV module input, 1 RV image represents the RV module input, $N = 4$ residual images represent the residual module input. Lastly, one camera image is used as input to the *LiCaNext* camera module. The color-coding in the BEV images is embraced for illustration purposes only, and they are the same as the labels image in Fig. 1.

The BEV, RV, and residual modules consist of two 3×3 convolution layers. In contrast, the camera module involves a lightweight pretrained network, followed by projecting and warping the features into RV form, and lastly, two 3×3 convolution layers. In order to project and warp camera features onto the RV form to be concatenated with the RV images, we first need to compute the mapping between the LIDAR points and the camera image. However, due to the different operating frequencies of LIDAR and camera sensors,

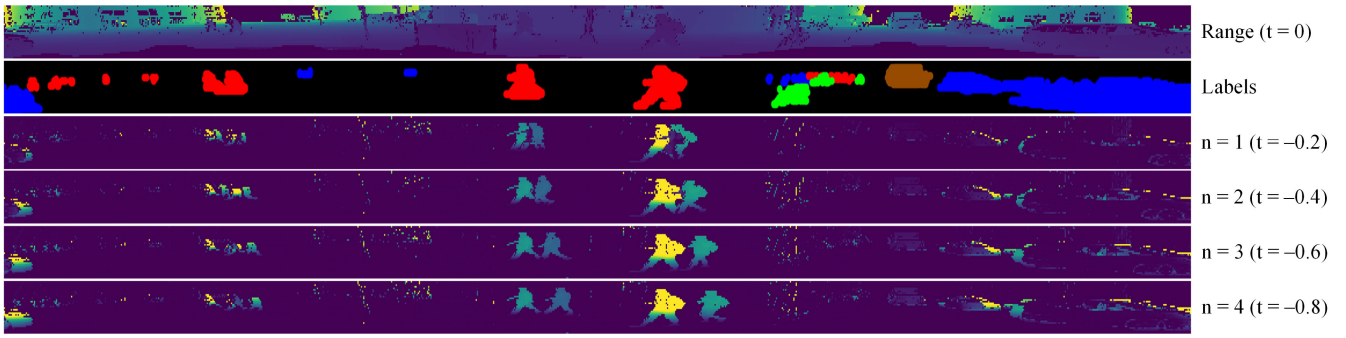

FIGURE 1. An illustrative example of normalized range residual images. The first two images portray the range of the current frame and the labels of all objects in that frame. A maximum of $N = 4$ residual images is demonstrated, with n indicating the residual image engendered between the previous n^{th} transformed frame and the current frame.

coordinate transformation must be applied to compute the mapping. The first step is to transform the LIDAR points into the vehicle's coordinate system at LIDAR capture time. Secondly, transform the points that exist in the vehicle's coordinate from LIDAR capture time to camera capture time. Next, transform the resulting points from the vehicle's coordinate system at camera capture time into the camera's coordinate system. Lastly, apply the camera's intrinsic calibration matrix on the transformed points. This transformation process compensates for the time shift between the two sensors and results in the LIDAR points being mapped correctly on the camera image. Suppose the features extracted from the pre-trained network have different dimensions than the original camera image. In that case, the mapping needs to be updated using a scale factor calculated between the dimensions of the extracted features and the original camera image. This mapping allows us to warp features from the camera image into RV representation.

The generated features from the RV, residual, and camera modules are then concatenated and sent into a U-net. The U-net consists of two scaling layers with a horizontal scaling factor of 2 on each layer. The vertical scale is kept constant due to its small representation compared to the width of the RV image. At each U-net layer, residual blocks with skip connections are used. The subsequent step is to project the resulting features from the U-net onto the BEV form to be concatenated with the BEV features encoded by the BEV module. Projection from one form to another is achieved using the painting approach. Algorithm 1 describes the projection from RV to BEV representation; however, the same approach is used to project from camera to RV representation. Basically, for each raw LIDAR point mapped to a BEV cell, its corresponding RV feature is projected into the same BEV cell position. If more than one RV feature ends up in the same cell, then the average is computed. Empty cells are filled with a value of -1. Further details on the projection algorithm can be found in [1].

The final step in *LiCaNext* multi-modal fusion network is to inject the concatenated features, in BEV form, into a single 3×3 convolution layer. At this stage, the produced features

Algorithm 1: Projecting Features From RV Into BEV

Inputs:
Lidar sweep: $\mathrm{L} \in \mathbb{R}^{N,\,D}$ with N points and $D = 4$.
RV features: $\mathrm{RV} \in \mathbb{R}^{W,\,H,\,C}$ with W width, H height, and C channels.
BEV dimensions: $bev_{dim} \in \mathbb{R}^2$.
Output:
Projected RV features: $\mathrm{P} \in \mathbb{R}^{bev_{dim}[0],\,bev_{dim}[1],\,C}$.

```
count = 0                                   // count ∈ R^{bev_dim[0], bev_dim[1]}
for l in L do
    l_BEV = project(l_{x,y})                // l_BEV ∈ R^2
    if l_BEV falls within bev_dim range then
        l_RV = project(l_{x,y})             // l_RV ∈ R^2
        tmp = RV[l_RV[0], l_RV[1],:]        // tmp ∈ R^C
        P[l_BEV[0], l_BEV[1],:] += tmp
        count[l_BEV[0], l_BEV[1],:] += 1
    end
    P /= count                              // average (avoid division by 0)
    mask = (count == 0)
    P[mask,:] = −1                          // assign −1 to empty cells
end
```

are rich and comprehensive. They consist of spatio-temporal information sourced from two representations (BEV and RV), physical object dimensions encoded in the input BEV images, occlusion information provided from RV images, and rich semantics signified in a camera image. When these features are inserted into MotionNet backbone network, they yield accurate pixel-wise joint perception and motion prediction in real-time.

C. MotionNet BACKBONE

MotionNet [4] is a novel model dedicated to performing pixel-wise joint perception and motion prediction in real-time for autonomous driving. The backbone network used in MotionNet is called spatio-temporal pyramid network (STPN). A spatio-temporal convolution (STC) block is the main element of STPN, which consists of two 2D

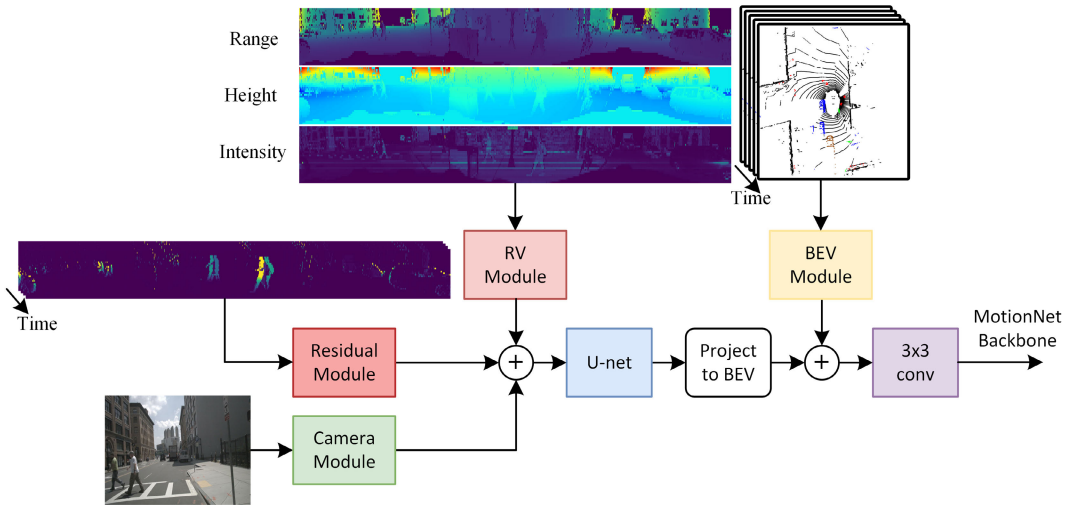

FIGURE 2. ***LiCaNext*** **architecture.**

convolutions followed by one pseudo-1D convolution. The STPN architecture has STC blocks structured in a hierarchical way to extract features at different scales, leveraging multi-scale spatio-temporal features. The lightweight design of the STC block is the key reason behind MotionNet speed efficiency, enabling it to run in real-time. The loss function of MotionNet is decomposed into six components, each of which is responsible for global or local regularization of the training network. The loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{motion} + \mathcal{L}_{state} + \alpha\ \mathcal{L}_s + \beta\ \mathcal{L}_{ft} + \gamma\ \mathcal{L}_{bt}, \quad (2)$$

The global regularization loss components ($\mathcal{L}_{class}$, $\mathcal{L}_{motion}$, and $\mathcal{L}_{state}$) are devoted for the three MotionNet output heads: cell classification, motion prediction, and state estimation, respectively. Cell classification output head predicts the category of each pixel, while motion prediction predicts their respective motion. Besides, the state estimation output head aims to predict whether each pixel is static or dynamic. We refer the reader to [4] for an illustrative figure describing the MotionNet architecture and the structure of its final output. $\mathcal{L}_{class}$ and $\mathcal{L}_{state}$ are cross-entropy losses, while $\mathcal{L}_{motion}$ is smooth L_1 loss. In cross-entropy, a distinct weight is assigned to each category to handle the class imbalance issue.

On the other hand, the remaining loss components ($\mathcal{L}_s$, $\mathcal{L}_{ft}$, and $\mathcal{L}_{bt}$) are associated with local regularization. The constants α, β, and γ in Eq. (2) are balancing factors. $\mathcal{L}_s$ loss component, defined in Eq. (3), limits the predicted motion of all pixels relating to an object o_k in one frame. So, the overall motion of o_k should be reflected by all pixels belonging to the same object.

$$\mathcal{L}_s = \sum_k \sum_{(i,j),(i',j')\in o_k} \left\| X_{i,j}^{(\tau)} - X_{i',j'}^{(\tau)} \right\|, \quad (3)$$

where $X_{i,j}^{(\tau)} \in \mathbb{R}^2$ is the predicted motion at pixel (i, j) in time τ. Comparing all $X_{i,j}^{(\tau)}$ and $X_{i',j'}^{(\tau)}$ is computationally expensive, hence only adjacent pixels are considered. $||\cdot||$ is L_1 loss.

Unlike $\mathcal{L}_s$, which restricts motion spatially, $\mathcal{L}_{ft}$ in (4) constrains the predicted motion temporally for foreground objects between adjacent frames. This is achieved by assuming that no sudden changes in motion will occur between two consecutive frames.

$$\mathcal{L}_{ft} = \sum_k \left\| X_{o_k}^{(\tau)} - X_{o_k}^{(\tau+\Delta t)} \right\|, \quad (4)$$

where $X_{o_k}^{(\tau)} \in \mathbb{R}^2$ denotes the overall motion of object o_k, computed as follows: $X_{o_k}^{(\tau)} = \sum_{(i,j)\in o_k} X_{i,j}^{(\tau)}/M$, where M is the number of pixels representing o_k.

Lastly, $\mathcal{L}_{bt}$ defined in Eq. (5) confines the temporal loss for static background pixels.

$$\mathcal{L}_{bt} = \sum_{(i,j)\in X^{(\tau)} \bigcap T(\tilde{X}^{(\tau-\Delta t)})} \left\| X_{i,j}^{(\tau)} - T_{i,j}(\tilde{X}^{(\tau-\Delta t)}) \right\|, \quad (5)$$

where $X^{(\tau)}$ and $\tilde{X}^{(\tau)}$ are motion predictions with current time being t and $t + \Delta t$, respectively. A transformation $T \in SE(3)$ is necessary to align $\tilde{X}^{(\tau-\Delta t)}$ with $X^{(\tau)}$. After applying T, the static background pixels in $T(\tilde{X}^{(\tau-\Delta t)})$ and $X^{(\tau)}$ will partially overlap.

IV. EXPERIMENTAL EVALUATION

A. DATASET

We use the public nuScenes dataset [16] to evaluate the proposed *LiCaNext* model extensively. The nuScenes dataset consists of 850 scenes, where each scene is a continuous sequence of length 40s. To perform our experiments, we partition the entire available scenes: 500 for training, 100 for validation, and the rest for testing. The dataset offers various sensors, including LIDAR, cameras, RADARs, GPS, and IMU. From this broad set of sensors, we only employ the LIDAR and the front camera sensors. The capture frequency

of LIDAR is 20Hz, and 12Hz for the camera sensor. The LIDAR has a complete horizontal field-of-view (FOV) and a vertical FOV ranging from $-30.67°$ to $10.67°$. Lastly, the camera has an opening angle of $70°$.

B. EXPERIMENTS

Table 1 comprises all experiments needed to verify the performance enhancement attained by *LiCaNext* in both perception and motion prediction. The first experiment records the performance of the original MotionNet model, which acts as the primary baseline. LiCaNet [1] proved that engaging RV and camera images into the fusion process outperform the baseline, which depends merely on BEV images as input. Our proposed *LiCaNext* expands on LiCaNet by incorporating residual images pushing the performance even further. To observe this performance advancement, we include two of the most prominent LiCaNet experiments in Table 1. The first LiCaNet experiment uses only a LIDAR sensor in the multi-modal fusion process, i.e., the fusion of BEV and RV images. The second LiCaNet experiment embraces LIDAR and camera sensors, i.e., the fusion of a camera image on top of BEV and RV images. The use of VGG16 in LiCaNet as the lightweight pretrained network for extracting high-level camera features achieved the best overall performance among the other evaluated pretrained networks. Accordingly, for *LiCaNext* experiments, we also adopt 6 layers of VGG16 for the lightweight pretrained network.

Before we evaluate the performance of the entire *LiCaNext* fusion model, we first examine the effect of fusing different numbers of range residual images on a multi-modal fusion network that depends solely on a LIDAR sensor. Therefore, the first series of *LiCaNext* (LIDAR only) experiments will omit the camera module and fuse only BEV, RV, and residual images. The outcome of this evaluation will be compared to LiCaNet (LIDAR only) experiment. Consequently, the number of sequential range residuals that realize the best performance is adopted to evaluate the entire *LiCaNext* model, including the camera module. In all experiments, the RV and range residual images are defined to have width and length dimensions of 1024×32, respectively.

C. EVALUATION METRICS

The metrics used in Table 1 to evaluate the success of the experiments are categorized into two groups: displacement error and classification accuracy. The displacement error measures the correctness of motion prediction, while the classification accuracy accesses the perception level. Each of these two metric groups is further divided into subgroups. The displacement error is divided into three-speed subgroups: static, slow, and fast, with mean and median errors computed for each. The static subgroup is defined for pixels with no motion, while the slow and fast subgroups are defined for pixels having speeds of $(0m/s, 5m/s]$ and $(5m/s, 20m/s]$, respectively. Thus, motion is predicted in a sequence of 20 frames, translating into 1s into the future. Only 5 predicted frames are selected for evaluation, with a 0.2s period between the sampled frames.

Pixels are categorized into five object classes: background, vehicle, pedestrian, bike, and others. The 'others' class is assigned to any detected object not classified in any of the first four classes. In addition to the classification accuracy of each object class, the mean classification accuracy (MCA) and overall accuracy (OA) are also computed under the classification group. MCA in (6) denotes the average classification accuracy of the five object classes, while OA in (7) is the average classification accuracy over all pixels. Furthermore, the inference time for each experiment is measured.

$$MCA = \frac{1}{M}\sum_{i=1}^{M}\frac{(Total\ \#\ of\ \ correct\ predictions)_{C_i}}{(Total\ \#\ of\ \ ground\ truths)_{C_i}}, \quad (6)$$

$$OA = \frac{1}{K}\sum_{k=1}^{K}\frac{(\#\ of\ \ correct\ predictions)_k}{N_k}, \quad (7)$$

where C_i denotes the category class i, while M is the total number of category classes; N refers to the number of pixels in image k, and K is the total number of images in a dataset. The total number of predictions or ground truths for C_i are measured using all images in a dataset K. Only non-empty pixels are considered for evaluation.

D. TRAINING SETUP

Following [1], the learning rate is initialized at 1.6×10^{-3} and terminated at 0.8×10^{-3}, with a decay factor of 0.5 every 10 epochs. The time span between consecutive LIDAR sweeps used to construct the historical BEVs and the sequential residual range images is 0.2s. The current sweep denoted by the keyframe is sampled at 2Hz for training and 1Hz for testing. The batch size used to train *LiCaNext* is 4.

E. RESULTS

Table 1 reveals that *LiCaNext* achieves outstanding joint perception and motion prediction results compared to its predecessor LiCaNet. We begin our evaluation by comparing *LiCaNext* (LIDAR only, r = 1) experiment to LiCaNet (LIDAR only). In this comparison, we investigate the effect of fusing one range residual image with BEV and RV images. It is worth noting that when the residual module is removed from *LiCaNext*, i.e., no range residual images are fused, its architecture becomes the same as LiCaNet. *LiCaNext* (LIDAR only, r = 1) experiment obtains a maximum gain of 1.2% and 0.2% in MCA and OA, respectively. Moreover, a drop in displacement error is recorded for all speed groups, indicating better motion prediction. This evinces that the fusion of a single residual image with BEV and RV images significantly advances motion prediction and offers a substantial rise in perception.

Next, we investigate the effect of increasing the number of fused residual images on performance. Table 1 unveils that an increase in the number of fused residual images establishes a monotonic rise in perception accuracy and a

TABLE 1. Perception and motion prediction comparison between our proposed *LiCaNext* and LiCaNet models. Performance of the original MotionNet model is also included. Pixels are assigned to static, slow, and fast speed groups if their predicted motion is $0m/s$, ($0m/s$, $5m/s$], and ($5m/s$, $20m/s$], respectively.

Method	Static		Slow		Fast		Classification Accuracy (%)							Time
	Mean	Median	Mean	Median	Mean	Median	Bg	Vehicle	Ped.	Bike	Others	MCA	OA	(ms)
MotionNet [4]	0.0236	0	0.2534	**0.0959**	1.0778	0.7346	97.5	91.2	74.9	22.1	65.9	70.3	96.3	**20.5**
LiCaNet (LIDAR only) [1]	0.0230	0	0.2531	0.0964	1.0547	0.7305	97.6	92.0	80.6	22.6	69.2	72.4	96.6	28.1
LiCaNext (LIDAR only, r = 1)	0.0229	0	0.2526	0.0960	1.0325	0.7256	97.9	92.6	82.6	24.1	71.0	73.6	96.8	29.2
LiCaNext (LIDAR only, r = 2)	0.0226	0	0.2499	0.0960	1.0231	0.7142	97.9	92.9	82.7	23.9	71.4	73.8	96.8	29.4
LiCaNext (LIDAR only, r = 3)	0.0223	0	0.2484	0.0960	1.0075	0.7073	97.9	93.1	82.5	23.7	72.7	74.0	96.8	29.5
LiCaNext (LIDAR only, r = 4)	**0.0222**	0	**0.2479**	0.0960	**0.9810**	**0.6866**	98.0	93.0	82.8	25.7	70.3	**74.0**	**96.9**	29.7
LiCaNet (VGG16_6) [1]	0.0224	0	0.2527	0.0969	1.0456	0.7300	97.8	92.4	84.0	23.2	71.9	73.9	96.9	31.0
LiCaNext (VGG16_6, r = 4)	**0.0221**	0	**0.2425**	0.0961	**0.9801**	**0.6842**	98.1	93.1	83.9	24.6	72.0	**74.3**	**96.9**	31.9

constant reduction in displacement error. The optimal performance is registered for *LiCaNext* (LIDAR only, r = 4) experiment, where 4 residual images are exploited in the multi-modal fusion process. A peak gain of 1.6% and 0.3% is established in MCA and OA, respectively. Furthermore, a 73.7mm and 43.9mm decrease in the mean and median errors for the fast-speed group is recorded, and a drop of 5.2mm and 0.4mm for the slow-speed. The mean error for the static group procured a reduction of 0.8mm with no median error. On the other hand, the *LiCaNext* (LIDAR only, r = 1) achieves a mean and median drop of 22.2mm and 4.9mm for the fast-speed group, 0.5mm and 0.4mm for slow-speed, and 0.1mm and 0m for the static group. This confirms that increasing the number of sequential residual images in the fusion process lowers the displacement error significantly.

Moreover, during the training stage, the model learns associations between the motion patterns embedded in the residual images and their corresponding objects presented in the other input representations. These learned associations positively influence perception. For instance, vehicles typically have higher speeds and are characterized by more pixels than pedestrians and bicyclists. So when strong-motion cues sourced from multiple neighboring pixels are fed into the model, this increases the model's confidence in categorizing those pixels as a vehicle. Thus, fusing motion cues in the form of range residuals enhance perception. Lastly, comparing the performance of *LiCaNext* (LIDAR only, r = 4) with the original MotionNet results in a significant gain in motion prediction across all speed groups, with a rise of 3.7% and 0.6% in MCA and OA, respectively. Even though the median error in the slow group is not the lowest for *LiCaNext* experiments; however, the obtained error is lower than what LiCaNet achieved.

The reason behind incorporating a maximum of 4 range residual images in *LiCaNext* is to match the number of fused BEV images generated from previous frames. After determining that fusing 4 residuals with BEV and RV images yields the best accuracy in joint perception and motion prediction, the subsequent step is to evaluate the performance of the entire *LiCaNext* model. Usually, the inclusion of camera images in the fusion process boosts performance because of the rich semantic information provided by the RGB images. The performance gain as a result of including a camera sensor is realized in LiCaNet [1], where LiCaNet (VGG16_6) experiment outperformed LiCaNet (LIDAR only). Therefore, integrating a camera image into the fusion of BEV, RV, and residual images should further push the performance. According to Table 1, *LiCaNext* (VGG16_6, r = 4) achieves a perception enhancement of 0.3% in MCA, and the percentage of OA is maintained, compared to *LiCaNext* (LIDAR only, r = 4). A noticeable drop is fulfilled in most speed groups, except the median error of the slow group, which increased just by 0.1mm. This confirms that adding a camera image onto the fusion of BEV, RV, and residual images improves perception and motion prediction even further. Comparing *LiCaNext* (VGG16_6, r = 4) to a model that fuses BEV, RV, and camera images (i.e., LiCaNet (VGG16_6)), we can see that 0.4% rise in MCA is registered, OA accuracy is maintained. In addition, a greater error drop is recorded in all speed groups for motion prediction. This reveals that incorporating residual images onto a multi-modal fusion network involving a camera module positively affects performance. Ultimately, the motion prediction advancement that *LiCaNext* accomplished compared to MotionNet is even better than what its LIDAR-only version procured. *LiCaNext* obtained an outstanding enhancement of 4.0% for MCA and 0.6% for OA compared to MotionNet.

Table 2 presents the effect of exploiting sequential range residuals on small and distant objects. Generally, small objects (e.g., pedestrians and bikes) have lower perception accuracy than larger objects (e.g., vehicles) as they are represented by much fewer pixels. Similarly, the accuracy drops naturally with increasing distance from the sensor. This is because the sensor's performance degrades progressively at capturing fine information with a farther distance. However, we unveil in Table 2 that *LiCaNext*, compared to LiCaNet, improves accuracy even further for small and distant objects. The perception accuracies presented are measured according to three distance range groups: short (S), medium (M), and far (F). The short-range is defined to include all pixel detections within the (0m, 10m] range. While medium- and far-ranges are defined for pixels in the (10m, 20m] and

TABLE 2. Evaluating the perception accuracies based on three distance ranges: Short (S) - (0m, 10m], medium (M) - (10m, 20m] and far (F) - (20m, 30m]. The results of the last two experiments are limited to the camera 70° FOV.

Method	Classification Accuracy (%)														
	Background			Vehicle			Pedestrian			Bike			Others		
	S	M	F	S	M	F	S	M	F	S	M	F	S	M	F
MotionNet [4]	98.3	97.3	95.8	94.5	91.0	81.8	77.4	74.5	73.3	29.1	19.5	17.0	76.1	62.6	52.5
LiCaNet (LIDAR only) [1]	98.4	97.5	96.1	94.3	92.1	84.8	83.8	78.7	78.1	30.9	18.0	**19.9**	78.5	65.7	58.5
LiCaNext (LIDAR only, r = 1)	98.5	97.7	96.6	94.7	93.0	85.1	86.5	**82.0**	80.0	31.4	21.4	18.3	79.7	69.1	60.4
LiCaNext (LIDAR only, r = 2)	98.5	97.6	95.6	95.0	93.1	85.8	86.6	81.6	80.2	31.2	21.1	18.4	80.2	70.3	60.6
LiCaNext (LIDAR only, r = 3)	98.5	97.6	96.5	**95.4**	**93.3**	86.1	**86.6**	81.7	80.0	30.9	20.5	19.2	**81.1**	**70.6**	**62.7**
LiCaNext (LIDAR only, r = 4)	**98.6**	**97.7**	**96.6**	95.1	93.2	**86.2**	86.5	81.6	**81.0**	**36.5**	**21.2**	18.6	78.6	68.3	60.3
LiCaNet (VGG16_6) [1]	98.7	97.2	95.3	95.2	94.6	87.5	90.3	86.7	84.0	35.2	23.6	35.0	85.5	79.5	67.7
LiCaNext (VGG16_6, r = 4)	**98.8**	**97.4**	**95.6**	**95.4**	**94.9**	**87.9**	**90.7**	**87.1**	**84.6**	**35.9**	**24.1**	**35.8**	**85.5**	**80.1**	**68.1**

(20m, 30m] ranges, respectively. Furthermore, the last two experiments involve a front camera sensor. To evaluate the effectiveness of the fused features, including the camera features, we restrict the perception measurements of the last two experiments to the camera 70° FOV. This is because the camera features are only incorporated in that area. Most of the improvements due to the camera features would be recognized in that same region. However, the perception accuracy of the other experiments is measured on the entire LIDAR 360° FOV.

To begin with, *LiCaNext* (LIDAR only, r = 4) compared to LiCaNet (LIDAR only) experiment achieves a 0.8% rise in accuracy for the vehicles category in the short-range group. On the other hand, pedestrians and bikes obtain a higher accuracy gain of 2.7% and 5.6% for the same group, respectively. This shows that the jump in detection rate for smaller objects is higher than bigger ones, proving that *LiCaNext* attains better accuracy for smaller objects. Next, we demonstrate how the fusion of residual images caused better improvements in perception for distant objects. The accuracy gain for vehicles and pedestrians in the far-range is 1.4% and 2.9%, respectively. The bikes category is the only exception where the accuracy dropped by 1.3%; however, the overall accuracy of bikes is still superior to LiCaNet (LIDAR only) by 3.1%. Hence, not only is the detection gain higher for most object classes in the far-range in relation to short-range, but also smaller objects (pedestrians) obtained a greater gain at distant objects than larger ones. This observation confirms that residual images assist in intensifying the detection rate for both small and distant objects.

Furthermore, comparing the drop in accuracy between the far- and short-range in both *LiCaNext* (LIDAR only, r = 4) and LiCaNet (LIDAR only), we notice the drop in accuracy is lower in most categories. For the background, vehicles, pedestrians, and others, *LiCaNext* (LIDAR only, r = 4) achieved a drop of 2%, 8.9%, 5.5%, and 18.3%; whereas, LiCaNet (LIDAR only) achieved a drop of 2.3%, 9.5%, 5.7%, and 20.0%, respectively. This shows that the fusion of residual images reduces the loss resulting from the natural degradation of sensors' performance with farther distances. A final observation is that experiments that fuse range residuals secured more significant gains compared to MotionNet than what LiCaNet accomplished, especially for small and distant objects.

After demonstrating, in Table 1, that exploiting residual images in a multi-fusion network involving a camera module enhances performance even further, we now further investigate the effect of that experiment on small and distant objects. The following comparisons will only be made between the last two experiments of Table 2, as their measurements are restricted to the camera 70° FOV. The last two experiments show that within the camera FOV, the perception accuracy of *LiCaNext* (VGG16_6, r = 4) outperforms LiCaNet (VGG16_6). The accuracy improvement attained in the short-range for vehicles is 0.2%, while 0.4% and 0.7% are acquired for pedestrians and bikes. The accuracy improvement for vehicles in the far-range is 0.4%; whereas, for pedestrians and bikes, it is 0.6% and 0.8%, respectively. This shows that the inclusion of residual images within the fusion process of BEV, RV, and camera images results in improved perception for all object categories, and even stronger accuracy is secured for small and distant objects. Moreover, the accuracy drop between the far- and short-range groups is lower in *LiCaNext* (VGG16_6, r = 4) compared to LiCaNet (VGG16_6). This indicates that the fusion of residual images onto BEV, RV, and camera images decreases even further the natural effect of performance degradation within the camera FOV at farther distances.

Expanding the LiCaNet (LIDAR only) network to include residual images has incurred an additional inference time of 1.6ms for *LiCaNext* experiments (without the camera module). The involvement of the camera module resulted in an increase of 2.2ms compared to *LiCaNext* (LIDAR only, r = 4). Furthermore, the fusion of BEV, RV, residual, and camera images resulted in 0.9ms increase compared to LiCaNet (VGG16_6). The total inference time of the entire *LiCaNext* model is 31.9ms, which is less than the real-time requirement (50ms). Overall, the provided results confirm that incorporating residual images has a significant effect on improving performance with a minimal increase in inference time.

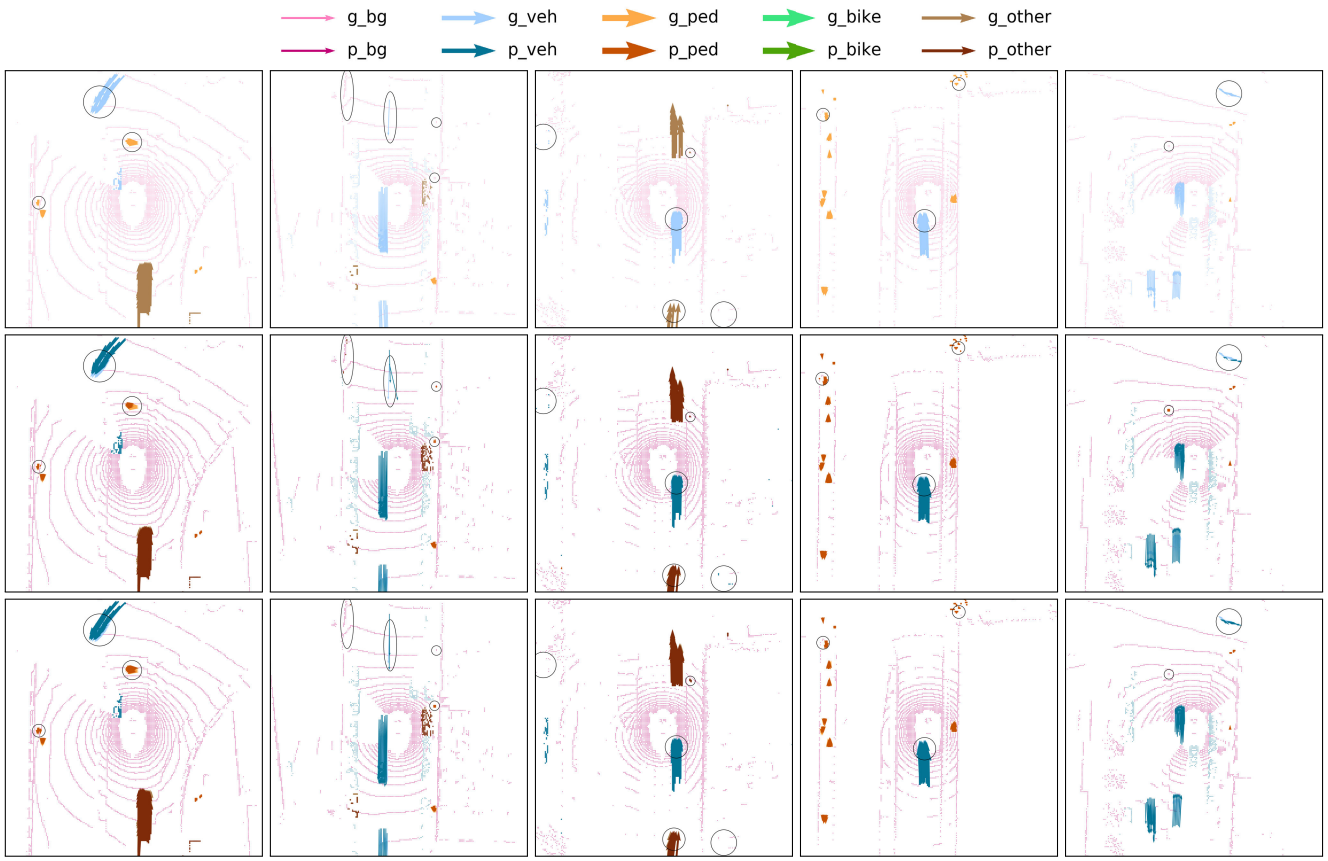

FIGURE 3. Comparison of perception and motion prediction using qualitative examples between *LiCaNext* (LIDAR only, r = 4) and LiCaNet (LIDAR only) in the full 360° range. The first row consists of only the ground truth. The second row displays the LiCaNet (LIDAR only) predictions, while the last row exhibits our *LiCaNext* (LIDAR only, r = 4) predictions. The ground truth is also present in the last two rows for easier visual comparison. The color codes used are attached at the top of the figure. *g_* denotes ground truth, while *p_* symbolizes prediction colors. *g_bg* and *p_bg*, for instance, represent the ground truth and predictions of the background pixels in the image, respectively.

Thus, *LiCaNext* can be used to perform accurate joint perception and motion prediction for autonomous driving.

Finally, we provide qualitative examples to illustrate the enhancements procured by *LiCaNext*. Fig. 3 compares the outcomes of *LiCaNext* (LIDAR only, r = 4) and LiCaNet (LIDAR only) using five different scenes. We identified the most noticeable gains with circles, but there are other modest improvements that *LiCaNext* achieves, particularly with motion; nevertheless, these are difficult to see. It is evident from the examples provided that *LiCaNext* results in more accurate perception and motion predictions. For instance, the motion predictions in the first example are more accurate in *LiCaNext* compared to LiCaNet, as the overlap between the motion predictions and the ground truth is higher in *LiCaNext*. The second example clearly shows how *LiCaNext* outperformed LiCaNet in terms of perception. The top right circle shows that LiCaNet mistakenly predicted several pixels as pedestrians, but *LiCaNext* correctly detected those pixels as background. Additionally, the middle right circle in the second example shows that both models mistakenly detected pedestrians even though there were none. The difference here is that *LiCaNext* predicted zero motion for those falsely detected pedestrians, whereas LiCaNet detected motion. Furthermore, the top right circle in the last example shows that *LiCaNext* predicted the vehicle's motion more accurately compared to LiCaNet. In that same example, LiCaNet wrongly predicted several pixels as pedestrians, whereas *LiCaNext* correctly predicted those as background.

To analyze the effect of fusing residual images in a multi-modal fusion network that involves a camera module, we provide three qualitative examples in Fig. 4 to compare between *LiCaNext* (VGG16_6, r = 4) and LiCaNet (VGG16_6). We only display the predictions within the vehicle's front view (70° FOV), as the camera features are only incorporated in that region, and most improvements will exist in that same region. Thus, restricting the view to the 70° FOV allows us to easily visualize the enhancements achieved due to adding residual images onto a multi-modal fusion network involving camera features. It is apparent from all three examples that the motion predictions attained by *LiCaNext* are closer to the ground truth compared to LiCaNet. The second example demonstrates how the perception of *LiCaNext* is better than LiCaNet. In that example, LiCaNet mistakenly detected a pixel as 'others'; however, *LiCaNext*

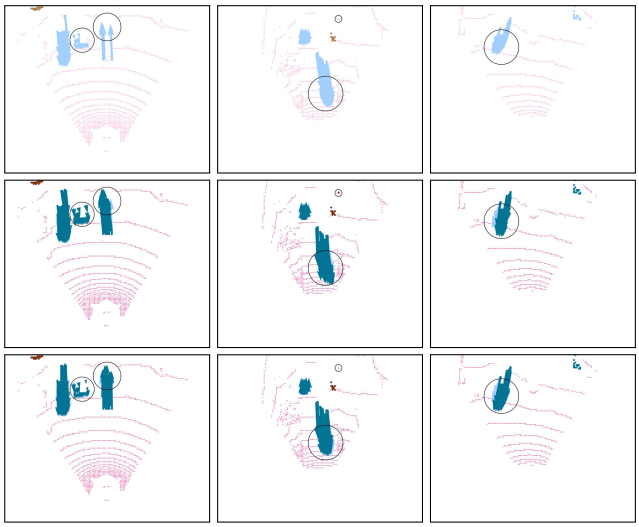

FIGURE 4. Qualitative comparison between the outcomes of *LiCaNext* (VGG16_6, r = 4) and LiCaNet (VGG16_6) within the camera 70° FOV. Rows indicate the ground truth, LiCaNet (VGG16_6), and *LiCaNext* (VGG16_6, r = 4) outcomes, respectively. The ground truth is also included in the last two rows.

correctly detected that pixel as background. Moreover, in the third example, we can see the overlap of *LiCaNext* motion predictions and the ground truth is greater than the overlap between the motion predictions obtained by LiCaNet and the ground truth. From the quantitative and qualitative results reported in this paper, we can confidently conclude that our proposed *LiCaNext* pushes the accuracy boundaries even further than LiCaNet.

V. CONCLUSION

In this paper, we put forward an accurate and real-time model, named *LiCaNext*, that performs pixel-wise joint perception and motion prediction. *LiCaNext* incorporates sequential range residual images in its multi-modal fusion process to significantly exceed the performance of its predecessor LiCaNet. *LiCaNext* performs exceptionally well on small and distant objects, achieving even better predictions than its previous version. Conducted experiments were evaluated on the challenging nuScenes dataset, confirming the excellent joint perception and motion prediction results.

REFERENCES

[1] Y. H. Khalil and H. T. Mouftah, "LiCaNet: Further enhancement of joint perception and motion prediction based on multi-modal fusion," *IEEE Open J. Intell. Transp. Syst.*, to be published.
[2] Y. H. Khalil and H. T. Mouftah, "End-to-end multi-view fusion for enhanced perception and motion prediction," in *Proc. 94th IEEE Veh. Techol. Conf.*, Nanjing, China, 2021, pp. 2.1–2.7.
[3] Y. H. Khalil and H. T. Mouftah, "Integration of motion prediction with end-to-end latent RL for self-driving vehicles," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Harbin, China, Jun. 2021, pp. 1111–1116.
[4] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11385–11395.
[5] S. Fadadu, S. Pandey, D. Hegde, Y. Shi, F.-C. Chou, N. Djuric, and C. Vallespi-Gonzalez, "Multi-view fusion of sensor data for improved perception and prediction in autonomous driving," 2020, *arXiv:2008.11901*. [Online]. Available: http://arxiv.org/abs/2008.11901
[6] M. Shah, Z. Huang, A. Laddha, M. Langford, B. Barber, S. Zhang, C. Vallespi-Gonzalez, and R. Urtasun, "LiRaNet: End-to-end trajectory prediction using spatio-temporal radar fusion," 2020, *arXiv:2010.00731*. [Online]. Available: http://arxiv.org/abs/2010.00731
[7] G. P. Meyer, J. Charland, S. Pandey, A. Laddha, S. Gautam, C. Vallespi-Gonzalez, and C. K. Wellington, "LaserFlow: Efficient and probabilistic object detection and motion forecasting," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 526–533, Apr. 2021.
[8] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss, "Moving object segmentation in 3D LiDAR data: A learning-based approach exploiting sequential data," 2021, *arXiv:2105.08971*. [Online]. Available: http://arxiv.org/abs/2105.08971
[9] G. Saur, W. Krüger, and A. Schumann, "Extended image differencing for change detection in UAV video mosaics," *Proc. SPIE*, vol. 9026, Mar. 2014, Art. no. 90260L.
[10] K. Sehairi, F. Chouireb, and J. Meunier, "Comparative study of motion detection methods for video surveillance systems," *J. Electron. Imag.*, vol. 26, no. 2, Apr. 2017, Art. no. 023025.
[11] L. S. P. Annabel and K. Sekaran, "Automatic signal clearance system using density based traffic control," in *Proc. 5th Int. Conf. Trends Electron. Informat. (ICOEI)*, Jun. 2021, pp. 1630–1635.
[12] M. Ramzy, H. Rashed, A. E. Sallab, and S. Yogamani, "RST-MODNet: Real-time spatio-temporal moving object detection for autonomous driving," 2019, *arXiv:1912.00438*. [Online]. Available: http://arxiv.org/abs/1912.00438
[13] H. Rashed, M. Ramzy, V. Vaquero, A. E. Sallab, G. Sistu, and S. Yogamani, "FuseMODNet: Real-time camera and LiDAR based moving object detection for robust low-light autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Oct. 2019, pp. 1–10.
[14] G. Rahmon, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Motion U-Net: Multi-cue encoder-decoder network for motion segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8125–8132.
[15] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and Bayesian GANs," *Neurocomputing*, vol. 394, pp. 178–200, Jun. 2020.
[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.
[17] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.
[18] J. Vertens, A. Valada, and W. Burgard, "SMSNet: Semantic motion segmentation using deep convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 582–589.
[19] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "MODNet: Motion and appearance based moving object detection network for autonomous driving," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2859–2864.
[20] A. Dewan, G. L. Oliveira, and W. Burgard, "Deep semantic classification for 3D LiDAR data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 3544–3549.
[21] E. Mohamed, M. Ewaisha, M. Siam, H. Rashed, S. Yogamani, W. Hamdy, M. Helmi, and A. El-Sallab, "Monocular instance motion segmentation for autonomous driving: KITTI instancemotseg dataset and multi-task baseline," 2020, *arXiv:2008.07008*. [Online]. Available: http://arxiv.org/abs/2008.07008
[22] H. Rashed, M. Essam, M. Mohamed, A. E. Sallab, and S. Yogamani, "BEV-MODNet: Monocular camera based bird's eye view moving object detection for autonomous driving," 2021, *arXiv:2107.04937*. [Online]. Available: http://arxiv.org/abs/2107.04937
[23] N. Djuric, H. Cui, Z. Su, S. Wu, H. Wang, F.-C. Chou, L. S. Martin, S. Feng, R. Hu, Y. Xu, A. Dayan, S. Zhang, B. C. Becker, G. P. Meyer, C. Vallespi-Gonzalez, and C. K. Wellington, "MultiXNet: Multiclass multistage multimodal motion prediction," 2020, *arXiv:2006.02000*. [Online]. Available: http://arxiv.org/abs/2006.02000
[24] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to predict intention from raw sensor data," in *Proc. Conf. Robot Learn.*, 2018, pp. 947–956.
[25] A. Laddha, S. Gautam, G. P. Meyer, C. Vallespi-Gonzalez, and C. K. Wellington, "RV-FuseNet: Range view based fusion of time-series LiDAR data for joint 3D object detection and motion forecasting," 2020, *arXiv:2005.10863*. [Online]. Available: http://arxiv.org/abs/2005.10863

[26] A. Laddha, S. Gautam, S. Palombo, S. Pandey, and C. Vallespi-Gonzalez, "MVFuseNet: Improving end-to-end object detection and motion forecasting through multi-view fusion of LiDAR data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2865–2874.
[27] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun, "PnPNet: End-to-end perception and prediction with tracking in the loop," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11553–11562.
[28] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3569–3577.
[29] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spatially-aware graph neural networks for relational behavior forecasting from sensor data," 2019, *arXiv:1910.08233*. [Online]. Available: http://arxiv.org/abs/1910.08233
[30] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "STINet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11346–11355.
[31] J. Phillips, J. Martinez, I. A. Bârsan, S. Casas, A. Sadat, and R. Urtasun, "Deep multi-task learning for joint localization, perception, and prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4679–4689.

YASSER H. KHALIL (Member, IEEE) received the B.Eng. degree (Hons.) in computer engineering from the American University of Kuwait, in 2015, and the M.Sc. degree (Hons.) from Kuwait University, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Ottawa. He has two years of industrial experience and over four years in academia. He is appointed as a Research Assistant at the University of Ottawa. His current research interests include artificial intelligence, deep learning, reinforcement learning, autonomous driving, sensor fusion, perception, and motion prediction.

HUSSEIN T. MOUFTAH (Life Fellow, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in computer science from the University of Alexandria, Egypt, in 1969 and 1972, respectively, and the Ph.D. degree in electrical engineering from Laval University, Canada, in 1975. He was with the ECE Department, Queen's University, from 1979 to 2002, where he was prior to his departure as a Full Professor and the Department Associate Head. He joined the School of Electrical Engineering and Computer Science (was the School of Information Technology and Engineering), University of Ottawa, in 2002, as a Tier 1 Canada Research Chair Professor, where he became a Distinguished University Professor, in 2006. He has six years of industrial experience mainly at Bell Northern Research of Ottawa (Nortel Networks). He is the author or the coauthor of 13 books, 73 book chapters, more than 1800 technical papers, 17 patents, six invention disclosures, and 148 industrial reports. He is the joint holder of 25 best/outstanding paper awards. He was a fellow the Canadian Academy of Engineering, in 2003, the Engineering Institute of Canada, in 2005, and the Royal Society of Canada RSC Academy of Science, in 2008. He has received numerous prestigious awards, such as the C. Gotlieb Medal in Computer Science and Engineering, the 2016 R. A. Fessenden Medal in Telecommunications Engineering of IEEE Canada, the 2015 IEEE Ottawa Section Outstanding Educator Award, the 2014 Engineering Institute of Canada K. Y. Lo Medal, the 2014 Technical Achievement Award of the IEEE Communications Society Technical Committee on Wireless Ad Hoc and Sensor Networks, the 2007 Royal Society of Canada Thomas W. Eadie Medal, the 2007–2008 University of Ottawa Award for Excellence in Research, the 2008 ORION Leadership Award of Merit, the 2006 IEEE Canada McNaughton Gold Medal, the 2006 EIC Julian Smith Medal, the 2004 IEEE ComSoc Edwin Howard Armstrong Achievement Award, the 2004 George S. Glinski Award for Excellence in Research of the University of Ottawa Faculty of Engineering, the 1989 Engineering Medal for Research and Development of the Association of Professional Engineers of Ontario, and the Ontario Distinguished Researcher Award of the Ontario Innovation Trust. He worked as the Editor-in-Chief of *IEEE Communications Magazine*, from 1995 to 1997, the IEEE ComSoc Director of Magazines, from 1998 to 1999, the Chair of the Awards Committee, from 2002 to 2003, the Director of Education, from 2006 to 2007, and a member of the Board of Governors, from 1997 to 1999 and from 2006 to 2007. He was a Distinguished Speaker of the IEEE Communications Society, from 2000 to 2007.