ETH zürich

ETH Library

HRFuser: A Multi-resolution Sensor Fusion Architecture for 2D Object Detection

Conference Paper

Author(s):
Brödermann, Tim; Sakaridis, Christos; Dai, Dengxin; Van Gool, Luc

Publication date:
2023

Permanent link:
https://doi.org/10.3929/ethz-b-000643408

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
https://doi.org/10.1109/ITSC57777.2023.10422432

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

HRFuser: A Multi-resolution Sensor Fusion Architecture for 2D Object Detection

Tim Brödermann[1], Christos Sakaridis[1], Dengxin Dai[2] and Luc Van Gool[1,3]

***Abstract*— Besides standard cameras, autonomous vehicles typically include multiple additional sensors, such as lidars and radars, which help acquire richer information for perceiving the content of the driving scene. While several recent works focus on fusing certain pairs of sensors—such as camera with lidar or radar—by using architectural components specific to the examined setting, a generic and modular sensor fusion architecture is missing from the literature. In this work, we propose HRFuser, a modular architecture for multi-modal 2D object detection. It fuses multiple sensors in a multi-resolution fashion and scales to an arbitrary number of input modalities. The design of HRFuser is based on state-of-the-art high-resolution networks for image-only dense prediction and incorporates a novel multi-window cross-attention block as the means to perform fusion of multiple modalities at multiple resolutions. We demonstrate via extensive experiments on nuScenes and the adverse conditions DENSE datasets that our model effectively leverages complementary features from additional modalities, substantially improving upon camera-only performance and consistently outperforming state-of-the-art 3D and 2D fusion methods evaluated on 2D object detection metrics. The source code is publicly available at https://github.com/timbroed/HRFuser**

I. Introduction

High-level visual perception is vital for the deployment of autonomous vehicles and robots. The primary sensors for such agents to perceive the surrounding scene are cameras, as they provide rich texture information at very high spatial resolution. This enables perception algorithms to achieve high accuracy in central tasks, such as object detection and semantic segmentation.

However, to attain full autonomy, systems require perception algorithms that perform robustly in all encountered conditions, but the quality of images degrades severely in adverse visual conditions, such as night-time, rainfall, snowfall, or fog. Moreover, camera readings do not explicitly capture depth or other geometric attributes of the scene. Complementary characteristics to cameras are provided by other sensors: lidars and radars provide explicit range measurements, while radars and gated cameras feature robustness to adverse weather [1]. Thanks to developments in sensor technology, these types of sensors are becoming cheaper and thus more commonly used in automated driving. Thus, exploiting *all* measurements from the sensor suite of an autonomous system via sensor fusion is of utmost importance for accurate perception under all possible conditions.

[1]Computer Vision Lab, ETH Zurich, 8092 Zurich, Switzerland {timbr,csakarid,vangool}@vision.ee.ethz.ch
[2]VAS, MPI for Informatics, 66123 Saarbrücken, Germany ddai@mpi-inf.mpg.de
[2]ESAT, KU Leuven, 3001 Leuven, Belgium

Besides cameras, adverse weather conditions can also severely affect the measurements of lidars [2], [3], [4]. This in turn results in lidar-based 3D object annotations being incomplete in such conditions. Fig. 1 displays both 2D and 3D labels from the DENSE dataset [1] and exemplifies why a significant amount of objects (41.91% for the "dense fog" split of DENSE) can receive only a 2D annotation when correct lidar measurements are missing due to environmental factors such as fog or precipitation. As these measurements do not provide a complete and reliable signal for creating 3D annotations. However, in difficult driving conditions, it is of utmost importance to detect *all* safety-relevant objects even if their precise localization in 3D is not possible.

We thus focus on 2D object detection, which allows to train and evaluate detection models not only on standard data, such as nuScenes [5], but also on extremely challenging data where 3D annotations are missing due to the factors mentioned above, such as DENSE [1]. We pursue this goal by building a modular architecture that treats the camera as the *primary* modality and adaptively fuses features from an arbitrary number of additional, *secondary* modalities in a modular and scalable manner.

Our network, named HRFuser, consists of a multi-resolution multi-sensor fusion architecture for 2D detection. The structure of HRFuser is based on the paradigm of preserving high-resolution representations throughout all layers of the backbone [6], [7]. We extend this architectural paradigm to multiple modalities and propose an *efficient* fusion design for our HRFuser, which scales well with the number of sensors. In particular, HRFuser includes parallel lightweight branches for each of the secondary input modalities. Solely the primary camera branch constructs additional high-dimensional lower-resolution features.

We repeatable fuse the sensors at multiple levels and at all resolutions of the camera branch. To facilitate this, we propose a novel multi-window cross-attention (MWCA) block. This block efficiently performs an attention-based fusion of the camera with each additional sensor in parallel, reducing the quadratic complexity of attention via multiple non-overlapping spatial windows. MWCA efficiently attends to the useful features of each sensor while ignoring noise, resulting in improved performance from *all* added sensors, even from radar, which is highly noisy.

Our architecture is generic and modular, as it handles all additional sensors in the same way, except for basic pre-processing. This allows leveraging multiple sensors, such as lidar, radar, and gated cameras, without the need to create specialized architectural components dedicated to each

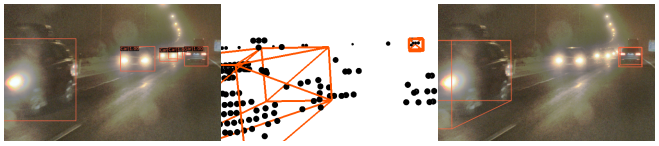

Fig. 1. An example scene from DENSE [1] with (left) 2D and (right) 3D object annotations. Multiple safety-relevant objects are missing in the (middle) point cloud due to weather deterioration and thus receive only 2D and not 3D annotations.

individual sensor. Thus, HRFuser is directly applicable to an arbitrary number of sensors. Our novel MWCA fusion and our architecture design with light-weight branches for secondary modalities minimize the computational overhead to only +9.7% flops and +1.9% parameters for a single added modality, as detailed in Sec. IV-C. HRFuser also inherits the benefits associated to processing camera features at multiple resolutions while preserving a high-resolution representation, allowing aggregation of global context without loss of fine spatial details.

We conduct a thorough experimental evaluation of our network for 2D detection on two major autonomous driving datasets, the adverse-condition-oriented DENSE [1] and the large-scale nuScenes [5]. HRFuser substantially outperforms all state-of-the-art 2D sensor fusion and camera-only networks which are heavily engineered for dense prediction tasks. As well as state-of-the-art 3D object detection approaches evaluated in 2D. Detailed ablation studies evidence the benefit of our carefully designed network architecture and the novel MWCA fusion block compared to other fusion strategies.

II. Related Work

Object detection methods output bounding boxes for a given input scene. A popular line of work on 2D detection consists in the region-based CNN (R-CNN) framework [8], [9], which employs a two-stage pipeline that first generates object proposals and then predicts the final boxes from the proposals. An alternative approach is single-stage detection [10], which is typically faster but less accurate. Recent approaches improve the efficiency of the detectors [11], and adapt networks to adverse conditions such as fog [12]. HRNet [6] constitutes a CNN backbone for detection that preserves a high resolution for intermediate representations, while aggregating global context via parallel lower-resolution branches. Recently, HRFormer [7] has extended this idea by replacing most convolutional blocks of HRNet with transformer blocks, which facilitate context aggregation via attending to features from any location of the input. HRFuser follows the architecture of HRNet and HRFormer with parallel streams of different resolutions, but it extends it to a multi-modal setting, by adding sensor-specific branches and a novel transformer-based fusion block, which allows us to simultaneously fuse information from multiple additional modalities in an adaptive manner.

Sensor fusion for object detection is the primary application of sensor fusion in visual perception, although other tasks [13], [14] have also been studied. For a comprehensive overview of related work, we refer the reader to [15]. The KITTI dataset [16] has catalyzed research in this area by providing recordings of driving scenes with multiple sensors, notably a lidar and a camera, along with object annotations. Successors of KITTI include nuScenes [5] and Argoverse [17]. Notably, nuScenes also includes radar readings, which are important in adverse-weather scenarios. Such scenarios are explicitly covered in [1], [18], [19]. Based on these sets, several sensor fusion works have been presented that focus on improving lidar-based 3D detection by fusing information from the camera. This category of works ranges from early (low-level) fusion [20], which directly combines the raw lidar data with raw image data or image features, and mid-level fusion [21], which combines lidar features with image-space features, to late fusion [22], which fuses the detection results from lidar and camera, asymmetric fusion [23], which fuses the object-level representations from one modality with data-level or feature-level representations from the other, to bird's-eye-view (BEV) based fusion [24], which lifts mid-level camera features to a common BEV space. Other sensor fusion methods address multi-modal 2D detection; many focus on radar and camera sensors [25], [26]. Methods that improve image-based 2D detection by fusing information only from lidar include [27]. Fewer previous works [1], [28], [29] fuse all three modalities, i.e. camera, lidar, and radar, for detection. We argue that using all three modalities is relevant, as they provide complementary characteristics which are essential for detection. While most recent works focus on fusing sensors with architectural components specific to individual sensors, we propose a modular fusion architecture for 2D object detection that easily scales to an arbitrary number of input modalities. Another feature of our network is the fusion at multiple levels and resolutions, which has also been applied in previous works [30], [1]. Different from these methods, our approach keeps high-resolution representations for each modality *throughout* the network in parallel with lower-resolution representations, which allows to better preserve details while also exploiting global context for classification.

Transformers [31] gained popularity in computer vision with the vision transformer [32]. More recent methods use local windows [7] and the Pyramid Vision Transformer (PVT) [33] introduces a spatial reduction attention to reduce the memory footprint. PVTv2 [34] improves upon PVT by using a linear-complexity attention module. To adaptively fuse two modalities with each other, cross-attention [35] was introduced. Different from these works, our transformer-based network handles several modalities instead of only two and fuses them at multiple resolutions, combining both global and local features of the input scene more effectively. Moreover, we combine local-window attention with cross-attention, thereby reducing the memory footprint and enabling repeated attention-based fusion at high resolutions.

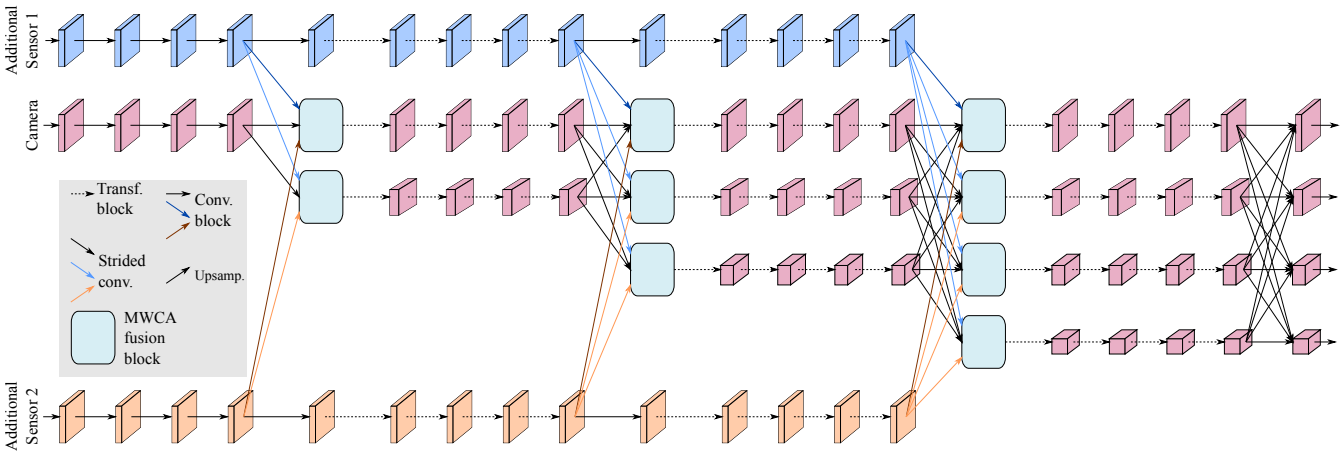

Fig. 2. An instantiation of the overall architecture of our HRFuser backbone for the case where two additional sensors besides the camera are available. Feature maps are colored according to the sensor branch to which they belong. For brevity, we only show the backbone of the network and not the detection head. Transf.: transformer, Conv.: convolution, MWCA: multi-window cross-attention.

III. HRFuser

With HRFuser, we extend the paradigm of preserving high-resolution representations throughout all layers of the backbone [6], [7] to multiple modalities. To this end, we extend the HRFormer [7] backbone with one additional high-resolution, but low-dimensional, branch for each added input modality besides the camera. These additional, or secondary, modalities are fused repeatedly at multiple resolutions into the branch of the primary modality.

Fig. 2 illustrates the general architecture of the multi-sensor fusion backbone of HRFuser. The design of the primary branch (camera) follows HRFormer, but is extended with a novel MWCA fusion block, which is further illustrated in Fig. 3. The MWCA fusion block is inserted between the multi-resolution fusion module and the subsequent transformer block, allowing the features from the secondary modalities to be fused into the camera branch. All secondary branches continue for three stages and are fused with the primary branch at three levels and four different resolutions. They include feature maps at a single, high resolution, while in the camera branch we introduce lower resolutions as we proceed to later stages, progressively aggregating context. Before applying our MWCA fusion blocks, we add 3×3 strided convolutions to match the high-resolution secondary modalities to the lower-resolution streams of the primary modality. This down-sampling causes the same 7×7 local window to progressively cover a larger area of the secondary modalities feature map. Our design, therefore, allows to keep detail in all modalities with the high-resolution stream, while efficiently fusing via local windows and still taking local and more global relationships into account.

Fusing multiple modalities in such an asymmetric way provides scalability to our method, as the complexity increases linearly with the number of added sensors. We can include an arbitrary number of modalities by simply adding an extra secondary branch for each new modality and fusing it in parallel into the camera branch. We demonstrate this possibility in Sec. IV by applying HRFuser to the DENSE dataset [1] and utilizing a gated camera as the fourth sensor, besides the more common lidar and radar sensors.

The HRFuser backbone illustrated in Fig. 2 is followed by a neck which forms a feature pyramid by concatenating the upsampled outputs of all streams [6]. This neck is in turn followed by a Cascade R-CNN head [9], following the widely used two-stage detector architecture. Cascade R-CNN introduces a sequence of detectors trained with increasing Intersection over Union (IoU) thresholds, setting a strong baseline for any given backbone.

Multi-window cross-attention. We propose a novel multi-window cross-attention (MWCA) block to fuse all modalities in parallel by applying multi-head cross-attention (CA) on multiple small non-overlapping local windows. In particular, MWCA limits the spatial extent of the cross-attention to small windows, addressing the quadratic complexity of attention and reducing the computational cost of each attention operation and allows to apply this operation to high-resolution feature maps. For each window, this results in K^2 tokens with dimensionality D, depending on the number of channels of the stream we fuse into. Compared to self-attention, CA fuses two modalities by applying attention with queries from the primary modality α and keys and values from the secondary modality β.

More formally, we partition the input feature map $\mathbf{X}$ of the primary modality α into a grid of P non-overlapping spatial windows: $\mathbf{X}^\alpha \xrightarrow{\text{Split}} \{\mathbf{X}_1^\alpha, \mathbf{X}_2^\alpha, \ldots, \mathbf{X}_P^\alpha\}$. Exactly the same partition is applied to the feature maps $\mathbf{Y}^\beta$ of all secondary modalities $\beta \in \{1, \ldots, M\}$: $\mathbf{Y}^\beta \xrightarrow{\text{Split}} \{\mathbf{Y}_1^\beta, \mathbf{Y}_2^\beta, \ldots, \mathbf{Y}_P^\beta\}$. All input feature maps are vectorized across the spatial dimensions and have the same shape $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times D}$, where N denotes the total number of spatial positions and D denotes the number of channels, and each window is of size $K\times K$.

A local transformer applies parallel CA to each corresponding set of windows independently. Parallel CA on the set of p-th windows is formulated as follows:

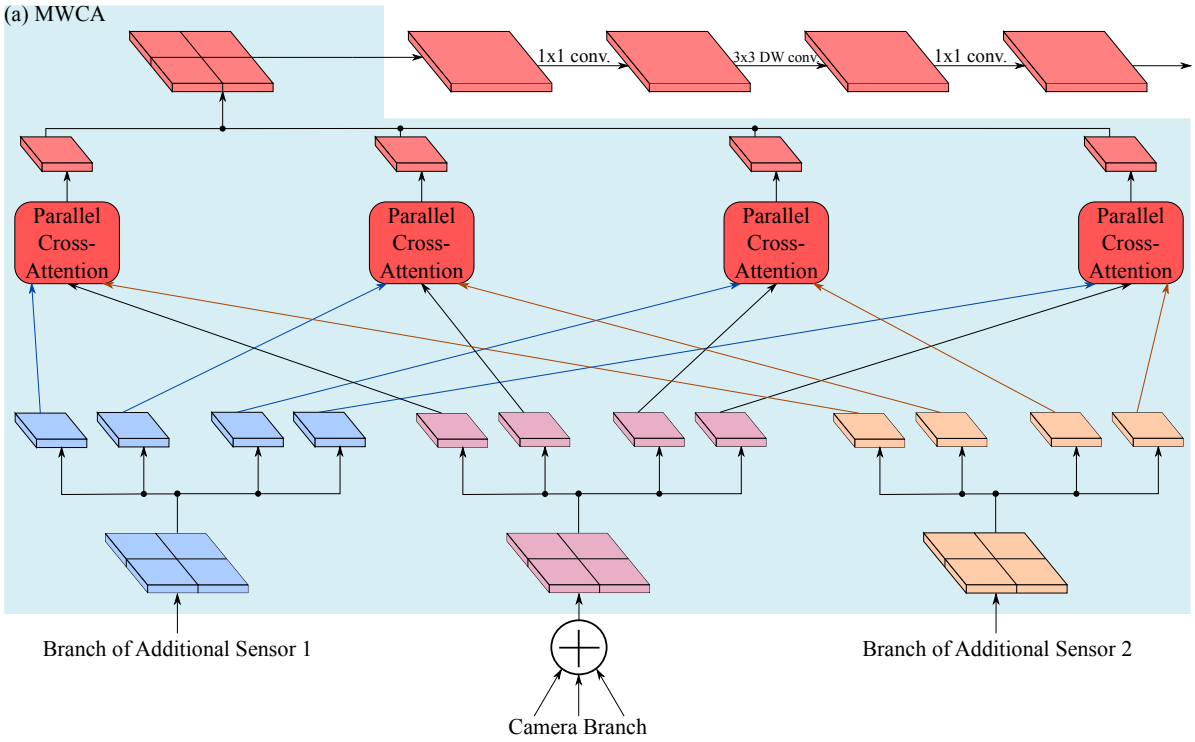

Fig. 3. Our multi-window cross-attention (MWCA) fusion block, consisting of (a) our MWCA and a subsequent feed-forward network. Inputs to the parallel cross-attention blocks are colored according to the sensor they come from. DW conv.: depth-wise convolution.

$$\begin{aligned}\text{MultiHead}(\mathbf{X}_p^\alpha, \mathbf{Y}_p^\beta) = \text{Concat}[&\text{head}(\mathbf{X}_p^\alpha, \mathbf{Y}_p^\beta)_1, \\ &\ldots, \\ &\text{head}(\mathbf{X}_p^\alpha, \mathbf{Y}_p^\beta)_H] \\ &\in \mathbb{R}^{K^2 \times D},\end{aligned} \tag{1}$$

$$\begin{aligned}\text{head}(\mathbf{X}_p^\alpha, \mathbf{Y}_p^\beta)_h = \text{Softmax}&\left[\frac{(\mathbf{X}_p^\alpha \mathbf{W}_q^{h,\beta})(\mathbf{Y}_p^\beta \mathbf{W}_k^{h,\beta})^T}{\sqrt{D/H}}\right] \\ &\mathbf{Y}_p^\beta \mathbf{W}_v^{h,\beta} \in \mathbb{R}^{K^2 \times \frac{D}{H}},\end{aligned} \tag{2}$$

$$\begin{aligned}\widehat{\mathbf{X}}_p = \mathbf{X}_p^\alpha + &\textstyle\sum_{\beta=1}^{M} \left[\mathbf{Y}_p^\beta + \text{MultiHead}(\mathbf{X}_p^\alpha, \mathbf{Y}_p^\beta)\mathbf{W}_o^\beta\right] \\ &\in \mathbb{R}^{K^2 \times D}\end{aligned} \tag{3}$$

where $\mathbf{W}_o^\beta \in \mathbb{R}^{D \times D}$ and $\mathbf{W}_q^{h,\beta}$, $\mathbf{W}_k^{h,\beta}$, $\mathbf{W}_v^{h,\beta} \in \mathbb{R}^{D \times \frac{D}{H}}$ for $h \in \{1, \ldots, H\}$ are weight matrices implemented by trainable linear projections. H denotes the number of heads and $\widehat{\mathbf{X}}_p$ denotes the output of the parallel CA for the set of p-th windows.

We arrange the outputs from all P sets of windows back into a single feature map to get the final output of MWCA, $\mathbf{X}^{\text{MWCA}}$:

$$\left\{\widehat{\mathbf{X}}_1, \widehat{\mathbf{X}}_2, \ldots, \widehat{\mathbf{X}}_P\right\} \xrightarrow{\text{Merge}} \mathbf{X}^{\text{MWCA}}. \tag{4}$$

Fig. 3 illustrates how we split up the input maps for each modality into non-overlapping windows and apply parallel CA across modalities within each window independently, before merging the resulting outputs back into a single feature map. Fig. 4 illustrates parallel CA in more detail. To allow information exchange between the non-overlapping windows, we add a feed-forward network including 3×3 depth-wise convolution.

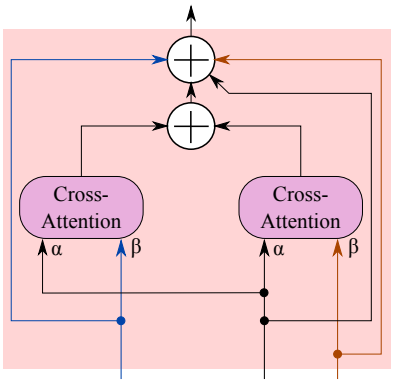

Fig. 4. Our parallel cross-attention block for the case where two additional sensors besides the camera are used. α denotes the primary modality (camera) and β denotes the secondary modalities.

Other architectural features. Before feeding inputs to HRFuser, we project all secondary modalities onto the image plane of the camera, using perspective projection as proposed in [36]. This yields an exact spatial correspondence between the input feature maps of different modalities, ensuring consistency among corresponding windows from different modalities in MWCA. All branches start with a CNN reducing the resolution by a factor of 4, followed by 4 stages consisting of multiple identical blocks. For all branches, we use basic bottleneck blocks to build the first stage [6] and transformer blocks to build all subsequent stages and streams [7]. We choose the parameters of each MWCA transformer (H, D) to be equal to the parameters of

TABLE I

COMPARISON OF 2D DETECTION METHODS ON NUSCENES EVALUATED ON 6 CLASSES FOLLOWING [26]. THE FIRST GROUP OF ROWS USES THE STANDARD NUSCENES SET, WHILE THE SECOND GROUP USES THE SPLITS FROM [25]. C: CAMERA, R: RADAR, L: LIDAR, (*): RESULTS TAKEN DIRECTLY FROM THE RESPECTIVE PAPER.

Method	Modalities	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
HRNetV2p-w18 [6]	C	32.4	56.6	33.5	21.0	43.7	43.4
HRFormer-T [7]	C	34.3	59.6	35.6	23.2	45.5	43.9
HRFormer-B [7]	C	33.8	59.4	34.6	22.4	45.1	43.1
Radar-Camera Fusion[26]*	CR	35.6	60.5	37.4	-	-	42.1
HRFuser-T	CRL	38.3	65.3	40.1	26.8	49.9	48.3
HRFuser-S	CRL	38.5	65.6	40.2	**27.2**	49.9	48.1
HRFuser-B	CRL	**38.8**	**66.0**	**41.0**	26.9	**50.7**	**48.6**
CRF-Net [25]	CR	27.0	42.7	29.0	22.7	35.6	31.3
HRFuser-T	CRL	**34.6**	**62.0**	**34.7**	**26.0**	**48.5**	**45.8**

the subsequent transformer blocks of the respective stream. We include additional implementation details on different versions of HRFuser in the supplement[1].

IV. EXPERIMENTS

We organize this section as follows. We first present our implementation details and experimental setup for multi-sensor 2D object detection on the two examined datasets, DENSE [1] and nuScenes [5]. We then compare the 2D performance of our method to the state-of-the-art in 2D and 3D multi-sensor fusion and conduct detailed ablation studies on the utility of including additional sensors and the fusion mechanism.

A. Experimental Setup

In all our experiments, we use a two-stage HRFuser network for 2D detection. The backbone of the network is structured as per Sec. III and its outputs are used to feed a Cascade R-CNN [9] head which serves as the second stage of the network. We test a tiny (T), small (S) and base (B) version of HRFuser and implement them using the mmdetection framework [37].

HRFuser is trained on DENSE for 60 epochs on batches of size 12 using AdamW with a base learning rate (LR) of 0.001. We apply a 500-step LR warm-up and reduce the LR by a factor of 10 at epochs 40 and 50. The training settings are the same for nuScenes, except that we use 12 epochs, a LR of 0.0001 and LR reductions at epochs 8 and 11. To accelerate learning of features from the less rich modalities such as radar, we randomly set inputs to zero during training [25], [1] with a chance of 50% for DENSE and 20% for nuScenes.

DENSE [1] is a multi-modal driving dataset with 106k 2D and 68k 3D bounding boxes. The dataset provides camera images, lidar and radar points, and gated camera images, captured under a variety of normal and adverse weather conditions. The gated camera in DENSE captures images in the NIR band at 808nm with a time-synchronized flood-lit flash laser source. Following the standard dataset splits in [1], we train only on clear-weather data and use adverse-condition data only for evaluation. We follow [1] for basic sensor pre-processing, obtaining 1248×360 images with depth, intensity and height for lidar, and depth and velocity over ground for radar. Note that radar is missing the RCS channel, since this is not published with the rest of DENSE. We train on the common KITTI classes car, pedestrian, and cyclist, and evaluate only on car using the KITTI evaluation framework [16], similar to [1].

NuScenes [5] is a large-scale dataset (1.4M images) providing 3D data and annotations of a full autonomous vehicle sensor suite including 6 cameras, 1 lidar and 5 radars. We follow [25] for basic sensor pre-processing, creating radar images with range, radar cross-section (RCS) and velocity over ground, and lidar images with range, intensity and height. Compared to [25], we do not accumulate radar data across time or filter them in any way. Unless otherwise stated, we use a subset of 10 nuScenes object classes following the mmdet3d [38] framework: car, truck, trailer, bus, construction vehicle, bicycle, motorcycle, pedestrian, traffic cone, and barrier. To create 2D ground truth and to evaluate 3D approaches in 2D, we project the 3D bounding boxes onto each image plane by computing a rectangle convex hull of the projected corners, similar to [25], [26]. Whereby, we discard annotations that are labeled with the lowest visibility bin, thereby filtering out occluded boxes. We train on the official training set and evaluate on the validation set, due to the lack of a public benchmark for 2D detection. Evaluation uses the 2D COCO evaluation metrics [39].

B. Comparison to the State of the Art

We compare multiple versions of HRFuser to state-of-the-art camera-only and multi-modal methods on nuScenes in Tab. I. All versions of HRFuser outperform substantially all camera-only models. In particular, the fully-fledged HRFuser-B improves AP by 5.0% compared to HRFormer-B and demonstrates analogous improvements in all other metrics. Moreover, all versions of HRFuser beat the radar-camera fusion method of [26] by a large margin on the standard nuScenes split, showcasing the advantage of leveraging multiple complementary sensors—including lidar—with a single, modular architecture as ours over just using radar and camera. A substantial performance gain of +7.6 AP is also observed over CRF-Net [25] on the nuScenes split that is employed by [25] and using only the front camera for evaluation.

In Tab. II, we compare our HRFuser-T to the fusion method of [1] on DENSE. Our model clearly outperforms [1] across all weather conditions, showing in particular significant improvements in the cases of light fog and dense fog, in which it beats [1] by 1.6% and 1.5% on the "hard" setting, respectively. This finding showcases the ability of our model to generalize well to previously unseen, adverse conditions, which degrade the quality of the readings for some of the sensors, such as the camera and the lidar, by

[1] https://github.com/timbroed/HRFuser/blob/master/Supplementary_Material_HRFuser.pdf

TABLE II

COMPARISON OF 2D DETECTION METHODS ON THE DENSE TEST SETS IN AP. (*): RESULTS TAKEN DIRECTLY FROM THE RESPECTIVE PAPER.

Weather	clear			light fog			dense fog			snow/rain		
Difficulty	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard
Deep Entropy Fusion [1]*	89.84	85.57	79.46	90.54	87.99	84.90	87.68	**81.49**	76.69	88.99	83.71	77.85
HRFuser-T	**90.15**	**87.10**	**79.48**	**90.60**	**89.34**	**86.50**	**87.93**	80.27	**78.21**	**90.05**	**85.35**	**78.09**

TABLE III

COMPARISON FOR 2D DETECTION ON NUSCENES. (’): 3D→2D PROJECTION IS USED TO OBTAIN 2D PREDICTIONS.

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
CenterPoint [40]’	17.9	40.9	13.1	8.1	28.2	29.8
BEVFusion [24]’	26.0	52.6	22.2	13.6	37.9	37.1
HRFuser-B	**32.9**	**58.9**	**33.0**	**23.8**	**44.1**	**43.5**

TABLE IV

ABLATIONS OF INPUT MODALITIES ON NUSCENES. RESULTS ARE IN AP. EF: EARLY FUSION, C: CAMERA, R: RADAR, L: LIDAR.

Modalities	C	CR	CL	CRL
HRFormer-T (EF)	26.5	25.7	**28.2**	27.7
HRFuser-T (ours)	26.5	27.9	31.2	**31.5**

properly attending to the features from the sensors that are more robust to these conditions, such as the radar and the gated camera.

In Tab. III, we compare our 2D HRFuser to 3D object detection approaches on nuScenes by projecting their 3D results to 2D, as indicated in Sec. IV-A. HRFuser substantially outperforms the state-of-the-art 3D detection method BEVFusion [24] on all 2D metrics, demonstrating its effectiveness for 2D detection against 3D-based approaches. Note that this comparison is reasonably fair as correct 3D predictions will still be correct when evaluated against the projected 2D ground truth.

C. Ablation Studies

Modalities. Tab. IV investigates the contribution of each sensor on nuScenes, by training HRFuser and a naive early fusion baseline with different subsets of input modalities. HRFormer-T (Early Fusion)—which naively utilizes a concatenated input without any additional changes to HRFormer—performs 1.7% better when adding lidar to the camera-only baseline. Note that the performance drops both times when we add the noisy radar to the input modalities. In contrast, adding radar to HRFuser yields an improvement of 1.4% over the camera-only baseline. The improvement is larger (4.7%) when adding lidar, and is maximized (5.0%) when combining all 3 sensors, showing the ability of our MWCA fusion to attend to the useful part of extra modalities—notably radar—while ignoring noisy content in them. HRFuser not only avoids a performance drop when adding radar but even gains additional performance. This result implies that the proposed method successfully pays attention to the relevant features.

We examine the effect of different modalities on DENSE in Tab. V. A combination of all four modalities yields the overall best performance, except for the case of dense fog, where a combination of camera, radar and gated camera performs best. This is in line with the findings of [1] and is due to the severe impact of fog on the lidar, as the laser pulse has to travel to the object and back, which squares the attenuation due to the presence of fog. By contrast, radar and gated cameras are more robust to fog. Note that the used standard splits of DENSE investigate the generalization capabilities rather than the robustness of a model since training includes only clear-weather data. Thus, the effect of fog on lidar is unseen during training, and the network cannot learn how to deal with the introduced noise, as it does with the radar noise on nuScenes in the previous paragraph. Another finding is that adding the gated camera on top of lidar and radar provides a consistent improvement across conditions, evidencing the informativeness of the high-resolution features from this sensor, which is generally robust to adverse conditions. Furthermore, we experimented with different sensors as primary modalities and found that choosing the information-dense RGB or gated cameras performed better than the sparse lidar and radar sensors. The higher spatial resolution may aid in guiding the fusion and attending to smaller details. For further details on the choice of primary modality, we refer the reader to the supplement.

Fusion mechanism. Tab. VI presents an ablation study on nuScenes regarding the fusion mechanism which is used in HRFuser, in order to verify the benefit of our MWCA fusion block. The reference is the camera-only HRFormer baseline. Early fusion achieves only a slight 1.2% improvement in AP over the camera-only HRFormer. Using our proposed HRFuser with its multi-resolution fusion design, but with a simplified addition-based fusion block instead of MWCA, already yields a large 4.3% improvement in AP over the camera-only baseline. Replacing addition with our proposed MWCA further improves performance consistently across all metrics, showcasing the utility of attention-based fusion for detection. Limiting the fusion to only the high-resolution stream of the camera branch yields a 1.0% reduction in AP, highlighting the importance of multi-resolution fusion. We compare our MWCA to an alternative attention mechanism via the state-of-the-art transformer PVTv2 [34], adapted for cross-attention (PVTv2-CA). For implementation details, we refer the reader to the supplement. Our MWCA fusion outperforms PVTv2-CA and the linear version PVTv2-Li-CA by 1.7% and 2.0% respectively, demonstrating the advantage

TABLE V

ABLATIONS OF INPUT MODALITIES FOR HRFUSER-T ON THE DENSE TEST SETS IN AP. C: RGB CAMERA, R: RADAR, L: LIDAR, G: GATED CAMERA.

C	L	R	G	clear			light fog			dense fog			snow/rain			Flops	Parameters	Inference
				easy	mod.	hard	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard	$[GFLOPs]$	$[M]$	$[ms]$
✓	×	×	×	79.81	62.48	53.68	80.84	63.07	62.08	71.84	62.69	54.05	78.68	61.19	52.72	104.0	47.9	81.1
✓	✓	×	×	89.91	85.16	78.68	90.47	88.44	80.55	87.39	78.32	71.13	89.21	79.88	76.19	114.1 (+9.7%)	48.8 (+1.9%)	103.3 (+27.4%)
✓	×	✓	×	88.48	80.15	76.25	90.37	86.40	79.6	88.51	79.72	71.87	88.13	78.85	70.27	114.0 (+9.6%)	48.8 (+1.9%)	103.2 (+27.3%)
✓	×	×	✓	89.76	85.37	78.36	90.56	88.04	80.47	88.67	80.64	72.25	89.62	80.14	76.58	114.0 (+9.6%)	48.8 (+1.9%)	103.0 (+27.0%)
✓	✓	✓	×	89.88	85.17	78.64	90.46	87.87	80.51	88.10	80.11	72.01	89.40	80.02	76.11	123.3 (+18.6%)	49.4 (+3.1%)	120.8 (+49.0%)
✓	✓	×	✓	90.14	**87.18**	79.44	90.62	89.17	80.95	88.56	80.33	72.21	**90.09**	85.32	**78.09**	123.2 (+18.5%)	49.4 (+3.1%)	121.0 (+49.2%)
✓	×	✓	✓	89.87	85.13	78.55	**90.64**	88.37	80.52	**88.97**	**80.86**	**78.64**	89.85	80.33	76.54	123.2 (+18.5%)	49.4 (+3.1%)	121.3 (+49.6%)
✓	✓	✓	✓	**90.15**	87.10	**79.48**	90.60	**89.34**	**86.50**	87.93	80.27	78.21	90.05	**85.35**	**78.09**	132.4 (+27.3%)	49.9 (+4.2%)	141.0 (+73.9%)

TABLE VI

ABLATIONS OF FUSION STRATEGIES ON NUSCENES.

Method (Fusion Type)	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
HRFormer-T	26.5	49.9	25.3	18.2	37.0	26.8
HRFormer-T (Early)	27.7	51.6	26.5	18.4	38.8	38.9
HRFuser-T (Addition)	30.8	56.4	30.5	22.0	41.9	42.0
HRFuser-T ($MWCA_{onlyHighRes}$)	30.5	56.1	29.7	21.8	41.4	41.5
HRFuser-T (MWCA)	**31.5**	**57.4**	**31.1**	**22.7**	**42.5**	**42.3**
HRFuser-T (PVTv2-CA [34])	29.8	54.3	29.4	20.1	41.3	40.9
HRFuser-T (PVTv2-Li-CA [34])	29.5	54.2	28.6	19.9	41.0	40.6

MWCA.

Efficiency. We further investigate the number of parameters/flops and the inference speed on an Nvidia Quadro RTX 6000 GPU in Tab. V. Our fusion method adds only a minor computational overhead. Even when using *all three* additional modalities besides the camera, the flops increase by only 27.3% and the parameter count by a marginal 4.2%. The inference time increases by 27.4% for one added modality and the full multi-modal network predicts in much less than double the time of the camera-only network.

D. Qualitative Results

The qualitative results on DENSE in Fig. 5 demonstrate that our proposed method is significantly more resilient to adverse conditions than a strong camera-only model. Note e.g. the second example, where HRFuser correctly detects obscured cars in the fog, while HRFormer misses them. Even though HRFuser misses a few distant objects in the other example, it still performs significantly better than HRFormer. HRFormer struggles particularly in detecting objects at a large distance. This can be attributed to the cumulative effect of atmospheric phenomena such as fog and snow on the appearance of objects as their distance from the camera increases. The good performance of HRFuser demonstrates its greater generalization capability thanks to learning robust features from multiple modalities.

Fig. 6 presents detection results on nuScenes of HRFormer, BEVFusion and HRFuser. HRFuser detects the partially occluded pedestrian in the first example, which is missed by HRFormer. The last example includes minimal queues from the camera. However, in contrast to the camera-only HRFormer and the lidar- and camera-based BEVFusion, HRFuser correctly detects both cars of the scene, showcasing its ability to effectively leverage complementary sensors for object detection.

V. CONCLUSION

We have proposed HRFuser, a multi-modal, multi-resolution and multi-level fusion architecture. In particular, we have extended the high-resolution paradigm for dense semantic prediction to multiple modalities by introducing additional high-resolution branches for the extra modalities besides the camera. HRFuser repeatedly fuses the extra modalities into the multi-resolution camera branch with a novel transformer block that applies cross attention in local windows and enables efficient learning of robust multi-modal features. Our proposed MWCA fusion module attends to discriminative information from additional sensors while ignoring their noisy parts. We have evaluated HRFuser on DENSE and nuScenes and demonstrated its state-of-the-art performance in 2D object detection across a wide range of scenes and conditions. Our architecture is generic and scales straightforwardly to an arbitrary number of sensors, thus being of particular relevance for practical multi-modal settings in autonomous cars and robots, which usually involve a diverse set of sensors.

ACKNOWLEDGMENT

This work was supported by the ETH Future Computing Laboratory (EFCL), financed by a donation from Huawei Technologies.

REFERENCES

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multi-modal sensor fusion in unseen adverse weather," in *CVPR*, 2020.

[2] R. H. Rasshofer, M. Spies, and H. Spies, "Influences of weather phenomena on automotive laser radar systems," *ARS*, 2011.

[3] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, "Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather," in *ICCV*, 2021.

[4] M. Hahner, C. Sakaridis, M. Bijelic, F. Heide, F. Yu, D. Dai, and L. Van Gool, "LiDAR snowfall simulation for robust 3D object detection," in *CVPR*, 2022.

[5] 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 *CVPR*, June 2020.

[6] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2021.

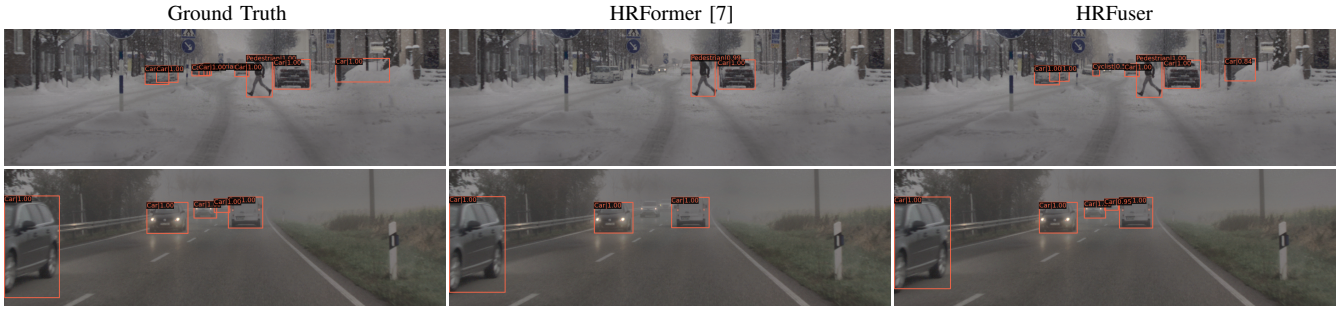

Fig. 5. Qualitative detection results on DENSE. Best viewed on a screen at full zoom.

Fig. 6. Qualitative detection results on nuScenes. Best viewed on a screen at full zoom.

[7] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," in *NeurIPS*, 2021.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[9] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *CVPR*, 2018.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.

[11] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *CVPR*, 2020.

[12] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *IJCV*, 2018.

[13] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *CVPR*, 2019.

[14] L. Wang, S. Giebenhain, C. Anklam, and B. Goldluecke, "Radar ghost target detection via multimodal transformers," *RA-L*, 2021.

[15] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *T-ITS*, 2021.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *CVPR*, 2012.

[17] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, June 2019.

[18] M. Pitropov, D. E. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander, "Canadian adverse driving conditions dataset," *IJRR*, 2020.

[19] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The Adverse Conditions Dataset with Correspondences for semantic driving scene understanding," in *ICCV*, 2021.

[20] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *CVPR*, 2020.

[21] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection," in *ECCV*, 2020.

[22] M. Roth, D. Jargot, and D. M. Gavrila, "Deep end-to-end 3D person detection from camera and lidar," in *ITSC*, 2019.

[23] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *CVPR*, 2017.

[24] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *ICRA*, 2023.

[25] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *SDF*, 2019.

[26] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv e-prints*, vol. abs/2009.08428, 2020.

[27] G. Melotti, C. Premebida, N. M. M. d. S. Goncalves, U. J. C. Nunes, and D. R. Faria, "Multimodal CNN pedestrian classification: A study on combining LIDAR and camera data," in *ITSC*, 2018.

[28] M. Pollach, F. Schiegg, and A. Knoll, "Low latency and low-level sensor fusion for automotive use-cases," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[29] S. S. Chaturvedi, L. Zhang, and X. Yuan, "Pay" attention" to adverse weather: Weather-aware attention-based object detection," in *ICPR*. IEEE, 2022.

[30] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *ECCV*, 2018.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021.

[34] ——, "Pvt v2: Improved baselines with pyramid vision transformer," *CVM*, 2022.

[35] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, 2019.

[36] Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li, and M. Tan, "Perception-aware multi-sensor fusion for 3d lidar semantic segmentation," in *ICCV*, 2021.

[37] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[38] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," https://github.com/open-mmlab/mmdetection3d, 2020.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[40] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *CVPR*, 2021.