



- (51) **International Patent Classification:**
G06K 9/00 (2006.01) *G06N 20/00* (2019.01)
- (21) **International Application Number:**
PCT/US2021/058615
- (22) **International Filing Date:**
09 November 2021 (09.11.2021)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
63/111,677 10 November 2020 (10.11.2020) US
63/113,945 15 November 2020 (15.11.2020) US
17/521,193 08 November 2021 (08.11.2021) US
- (71) **Applicant: NEC LABORATORIES AMERICA, INC.**
[US/US]; 4 Independence Way, Suite 200, Princeton, NJ
08540 (US).
- (72) **Inventors: LIU, Buyu;** 11527 Bianchini Lane, Cupertino, CA 95014 (US). **ZHUANG, Bingbing;** 250 Brandon Street, San Jose, CA 95134 (US). **CHANDRAKER, Manmohan;** 444 Saratoga Avenue, Santa Clara, CA 95050 (US).
- (74) **Agent: BITETTO, James, J.;** Tutunjian & Bitetto, P.C., 401 Broadhollow Road, Suite 402, Melville, NY 11747 (US).
- (81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) **Title:** END-TO-END PARAMETRIC ROAD LAYOUT PREDICTION WITH CHEAP SUPERVISION

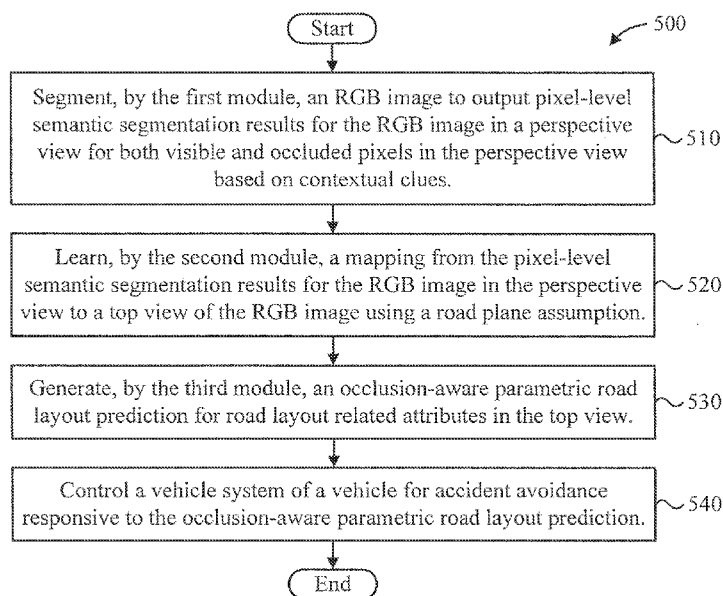


FIG. 5

(57) **Abstract:** A computer-implemented method for road layout prediction is provided. The method includes segmenting (510), by a first processor-based element, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues. The method further includes learning (520), by a second processor-based element, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption. The method also includes generating (530), by a third processor-based element, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

END-TO-END PARAMETRIC ROAD LAYOUT PREDICTION WITH CHEAP SUPERVISION

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Patent Application No. 17/521,193, filed on November 8, 2021, U.S. Provisional Patent Application No. 63/111,677, filed on November 10, 2020, and U.S. Provisional Patent Application No. 63/113,945, filed on November 15, 2020, incorporated herein by reference in their entireties.

BACKGROUND

Technical Field

[0002] The present invention relates to artificial intelligence and more particularly to end-to-end parametric road layout prediction with cheap supervision.

Description of the Related Art

[0003] Understanding road layout in top view with perspective input is very essential for real world applications such as autonomous driving or path planning. Recent work proposes to take an RGB image as input and provide pixel-level semantic predictions in top-view. However, such methods usually require pixel-level annotations, which can be very expensive. Moreover, their representations are not compact and do not have the ability to reason the occlusion relationship. Top-view representations can be more beneficial when occlusion relationships are desired, e.g., two objects cannot occupy the same position in top-view while they can potentially occlude each other in perspective view. Hence, an improved method for understanding road layout in top view is desired.

SUMMARY

[0004] According to aspects of the present invention, a computer-implemented method for road layout prediction is provided. The method includes segmenting, by a first processor-based element, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues. The method further includes learning, by a second processor-based element, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption. The method also includes generating, by a third processor-based element, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view.

[0005] According to other aspects of the present invention, a computer program product for road layout prediction is provided. The computer program product includes a non-transitory computer readable storage medium having program instructions embodied therewith. The program instructions are executable by a computer to cause the computer to perform a method. The method includes segmenting, by a first processor-based element of the computer, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues. The method further includes learning, by a second processor-based element of the computer, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption. The method also includes generating, by a third processor-based element of the computer, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view

[0006] According to yet other aspects of the present invention, a computer processing system for road layout prediction is provided. The computer processing system includes a memory device for storing program code. The computer processing system further includes a processor device for running the program code to segment an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues. The processor device further runs the program code to learn a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption. The processor device also runs the program code to generate an occlusion-aware parametric road layout prediction for road layout related attributes in the top view.

[0007] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0008] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0009] FIG. 1 is a block diagram showing an exemplary computing device, in accordance with an embodiment of the present invention;

[0010] FIG. 2 is a block diagram showing an exemplary end-to-end model, in accordance with an embodiment of the present invention;

[0011] FIG. 3 is a block diagram showing an exemplary system, in accordance with an embodiment of the present invention;

[0012] FIG. 4 is a block diagram showing an exemplary environment to which the present invention can be applied, in accordance with an embodiment of the present invention; and

[0013] FIG. 5 is a flow diagram showing an exemplary method for road layout prediction, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0014] Embodiments of the present invention are directed to end-to-end parametric road layout prediction with cheap supervision.

[0015] Embodiments of the present invention target end-to-end trainable top view layout estimation with a perspective image as input. There are two major challenges within the problem. First, the traditional method direct RGB method does not perform well on distance related attributes or on occluded regions. Second, existing SOTA methods have strong assumptions such as entire video sequences are given or have strong supervision requirements, e.g. pixel-wise semantic annotation or dense depth supervision. Thus, embodiments of the present invention propose to tackle those two above problems with a novel model that effectively exploits limited human annotations only in parametric form in top view, which is also able to output meaningful intermediate representations such as pixel-level occlusion-aware semantic segmentation in perspective view and semantics in top view.

[0016] Embodiments of the present invention propose an end-to-end model that first of all, outputs compact parametric predictions for road layout in top view with a perspective image as input. More importantly, the end-to-end model is able to output pixel-level semantic representations in both perspective view and top-view without corresponding human annotations. In an embodiment, the end-to-end model includes

three modules. The first module takes RGB as input and outputs semantic segmentation results in perspective view. The second module inputs the initial perspective semantics and learns to map them to the top-view. Finally, the third module receives the outputs from the second module and provides parametric predictions for road layout related attributes in top-view. The outputs of all three modules are occlusion-aware. To guide the training process, the present invention uses the unitize deep supervision that injects losses not only at the last module but also the first and second modules. Note that deep supervision in accordance with the present invention is “cheap” which means that human annotations are only needed at the last step, or in parametric form in top-view. Experimentally, it takes only seconds to annotate one image. Interestingly, with such cheap supervision in parametric form, the present invention is able to obtain a dense pixel-level supervision signal for module one and module two. Moreover, since the human annotations focus on only road layout, the present invention is even able to automatically recover layout related pixels, namely road, crosswalk, sidewalk or lane boundaries, that are originally occluded by foreground in perspective view without undue effort.

[0017] FIG. 1 is a block diagram showing an exemplary computing device 100, in accordance with an embodiment of the present invention. The computing device 100 is configured to perform road layout prediction.

[0018] The computing device 100 may be embodied as any type of computation or computer device capable of performing the functions described herein, including, without limitation, a computer, a server, a rack based server, a blade server, a workstation, a desktop computer, a laptop computer, a notebook computer, a tablet computer, a mobile computing device, a wearable computing device, a network appliance, a web appliance, a distributed computing system, a processor-based system,

and/or a consumer electronic device. Additionally or alternatively, the computing device 100 may be embodied as a one or more compute sleds, memory sleds, or other racks, sleds, computing chassis, or other components of a physically disaggregated computing device. As shown in FIG. 1, the computing device 100 illustratively includes the processor 110, an input/output subsystem 120, a memory 130, a data storage device 140, and a communication subsystem 150, and/or other components and devices commonly found in a server or similar computing device. Of course, the computing device 100 may include other or additional components, such as those commonly found in a server computer (e.g., various input/output devices), in other embodiments. Additionally, in some embodiments, one or more of the illustrative components may be incorporated in, or otherwise form a portion of, another component. For example, the memory 130, or portions thereof, may be incorporated in the processor 110 in some embodiments.

[0019] The processor 110 may be embodied as any type of processor capable of performing the functions described herein. The processor 110 may be embodied as a single processor, multiple processors, a Central Processing Unit(s) (CPU(s)), a Graphics Processing Unit(s) (GPU(s)), a single or multi-core processor(s), a digital signal processor(s), a microcontroller(s), or other processor(s) or processing/controlling circuit(s).

[0020] The memory 130 may be embodied as any type of volatile or non-volatile memory or data storage capable of performing the functions described herein. In operation, the memory 130 may store various data and software used during operation of the computing device 100, such as operating systems, applications, programs, libraries, and drivers. The memory 130 is communicatively coupled to the processor 110 via the I/O subsystem 120, which may be embodied as circuitry and/or components

to facilitate input/output operations with the processor 110, the memory 130, and other components of the computing device 100. For example, the I/O subsystem 120 may be embodied as, or otherwise include, memory controller hubs, input/output control hubs, platform controller hubs, integrated control circuitry, firmware devices, communication links (e.g., point-to-point links, bus links, wires, cables, light guides, printed circuit board traces, etc.) and/or other components and subsystems to facilitate the input/output operations. In some embodiments, the I/O subsystem 120 may form a portion of a system-on-a-chip (SOC) and be incorporated, along with the processor 110, the memory 130, and other components of the computing device 100, on a single integrated circuit chip.

[0021] The data storage device 140 may be embodied as any type of device or devices configured for short-term or long-term storage of data such as, for example, memory devices and circuits, memory cards, hard disk drives, solid state drives, or other data storage devices. The data storage device 140 can store program code for road layout prediction. The communication subsystem 150 of the computing device 100 may be embodied as any network interface controller or other communication circuit, device, or collection thereof, capable of enabling communications between the computing device 100 and other remote devices over a network. The communication subsystem 150 may be configured to use any one or more communication technology (e.g., wired or wireless communications) and associated protocols (e.g., Ethernet, InfiniBand®, Bluetooth®, Wi-Fi®, WiMAX, etc.) to affect such communication.

[0022] As shown, the computing device 100 may also include one or more peripheral devices 160. The peripheral devices 160 may include any number of additional input/output devices, interface devices, and/or other peripheral devices. For example, in some embodiments, the peripheral devices 160 may include a display, touch screen,

graphics circuitry, keyboard, mouse, speaker system, microphone, network interface, and/or other input/output devices, interface devices, and/or peripheral devices.

[0023] Of course, the computing device 100 may also include other elements (not shown), as readily contemplated by one of skill in the art, as well as omit certain elements. For example, various other input devices and/or output devices can be included in computing device 100, depending upon the particular implementation of the same, as readily understood by one of ordinary skill in the art. For example, various types of wireless and/or wired input and/or output devices can be used. Moreover, additional processors, controllers, memories, and so forth, in various configurations can also be utilized. These and other variations of the processing system 100 are readily contemplated by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0024] As employed herein, the term “hardware processor subsystem” or “hardware processor” can refer to a processor, memory (including RAM, cache(s), and so forth), software (including memory management software) or combinations thereof that cooperate to perform one or more specific tasks. In useful embodiments, the hardware processor subsystem can include one or more data processing elements (e.g., logic circuits, processing circuits, instruction execution devices, etc.). The one or more data processing elements can be included in a central processing unit, a graphics processing unit, and/or a separate processor- or computing element-based controller (e.g., logic gates, etc.). The hardware processor subsystem can include one or more on-board memories (e.g., caches, dedicated memory arrays, read only memory, etc.). In some embodiments, the hardware processor subsystem can include one or more memories that can be on or off board or that can be dedicated for use by the hardware processor subsystem (e.g., ROM, RAM, basic input/output system (BIOS), etc.).

[0025] In some embodiments, the hardware processor subsystem can include and execute one or more software elements. The one or more software elements can include an operating system and/or one or more applications and/or specific code to achieve a specified result.

[0026] In other embodiments, the hardware processor subsystem can include dedicated, specialized circuitry that performs one or more electronic processing functions to achieve a specified result. Such circuitry can include one or more application-specific integrated circuits (ASICs), FPGAs, and/or PLAs.

[0027] These and other variations of a hardware processor subsystem are also contemplated in accordance with embodiments of the present invention

[0028] FIG. 2 is a block diagram showing an exemplary end-to-end model 200, in accordance with an embodiment of the present invention

[0029] Basically, the intermediate outputs are 211 and 240. They are where deep supervisions are for. Three modules are 210, 260 (215 and 230) and 250. Among them, 215 and 230 are together for transforming and hallucinating the pixel-level perspective representations to top-view.

[0030] Thus, the end-to-end model 200 includes three modules. A first module 210, also referred to herein as a perspective predictor module or perspective module in short, predicts the semantic class of each pixel, including even unseen or occluded regions. A second module 260, formed from modules 215 and 230, also referred to herein as Top-view Semantics (TS) module, learns to map the perspective prediction to the top-view with plane assumption. A refinement module 230 learns to refine the unseen or noisy regions in initial bevinit (birds eye view initialization) obtained with module 220 and outputs refined bevinit 240. A third module 250, also referred to herein as a parametric predictor module or parametric module in short, provides the parametric representation

for a generated top-view map (the parametric representation from refined bevinit.). Note that human annotation is only obtained for the third module 250 and the supervisions for the first two modules 210 and 260 are obtained automatically. Compared to existing work, the end-to-end model 200 intellectually utilizes two intermediate representations without labor-intense and ambiguous annotations.

[0031] Hence, the present invention proposes an end-to-end model that first of all, outputs compact parametric predictions for road layout in top view with a perspective image as input. More importantly, the end-to-end model in accordance with the present invention is able to output pixel-level semantic representations in both perspective view and top-view without requiring corresponding human requiring human annotations.

[0032] As noted above, the end-to-end model 200 includes three modules. The first module 210 takes RGB 201 as input and outputs semantic segmentation results in perspective view. The second module 260 inputs the initial perspective semantics and learns to map the initial perspective semantics to top-view. Finally, the third module 250 receives the outputs from the second module 260 and provides parametric predictions for road layout related attributes in top-view. To guide the training process, the present invention proposes the unitize deep supervision that injects losses not only at the third module 230 but also the first module 210 and second module 260. Note that deep supervision in accordance with the present invention is also cheap which means that human annotation is only required at the last step, or in parametric form in top-view. Experimentally, it takes only seconds to annotate one image. Interestingly, with such cheap supervision in parametric form, the present invention is able to obtain dense pixel-level supervision signals for the first module 210 and the second module 220. Moreover, since the human annotations focus on only road layout, the present invention is even able to automatically recover layout related pixels, namely road, crosswalk,

sidewalk or lane boundaries, that are originally occluded by foreground in perspective view with minimal efforts.

[0033] FIG. 3 is a block diagram showing an exemplary system 300, in accordance with an embodiment of the present invention.

[0034] The system 300 receives an RGB image 301. The system 300 includes a perspective module 310 (first module), a feature transform module 320 (part of second module), a hallucination module 330 (part of second module), and a parametric prediction module 340 (third module). The feature transform module 320 and the hallucination module 330 form the TD module 220. The system 300 outputs (i) a pixel-level occlusion aware semantic prediction in perspective view 351 from the perspective module 310, (ii) a pixel-level semantics in top view 352 from the hallucination module 330, and (iii) a parametric attribute prediction in top view 353 from the parametric prediction module 340. These elements are described in further detail hereinbelow.

[0035] FIG. 4 is a block diagram showing an exemplary environment 400 to which the present invention can be applied, in accordance with an embodiment of the present invention.

[0036] In the environment 400, a user 488 is located in a scene with multiple objects 799, each having their own locations and trajectories. The user 488 is operating a vehicle 472 (e.g., a car, a truck, a motorcycle, etc.) having an ADAS 477.

[0037] The ADAS 477 calculates an occlusion-aware parametric road layout prediction.

[0038] Responsive to the occlusion-aware parametric road layout prediction, a vehicle controlling decision is made. To that end, the ADAS 477 can control, as an action corresponding to a decision, for example, but not limited to, steering, braking, and accelerating systems.

[0039] Thus, in an ADAS situation, steering, accelerating/braking, friction (or lack of friction), yaw rate, lighting (hazards, high beam flashing, etc.), tire pressure, turn signaling, and more can all be efficiently exploited in an optimized decision in accordance with the present invention.

[0040] The system of the present invention (e.g., system 400) may interface with the user through one or more systems of the vehicle 472 that the user is operating. For example, the system of the present invention can provide the user information through a system 472A (e.g., a display system, a speaker system, and/or some other system) of the vehicle 472. Moreover, the system of the present invention (e.g., system 400) may interface with the vehicle 472 itself (e.g., through one or more systems of the vehicle 472 including, but not limited to, a steering system, a braking system, an acceleration system, a steering system, a lighting (turn signals, headlamps) system, etc.) in order to control the vehicle and cause the vehicle 472 to perform one or more actions. In this way, the user or the vehicle 472 itself can navigate around these objects 499 to avoid potential collisions there between. The providing of information and/or the controlling of the vehicle can be considered actions that are determined in accordance with embodiments of the present invention.

[0041] FIG. 5 is a flow diagram showing an exemplary method 500 for road layout prediction, in accordance with an embodiment of the present invention.

[0042] At block 510, segment, by the first module, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues.

[0043] At block 520, learn, by the second module, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption. The pixel-level semantic

segmentation results are occlusion-reasoned to indicate items occluded in the perspective view of the RGB image. The top view of the RGB image is occlusion-reasoned to indicate items occluded in the perspective view of the RGB image.

[0044] At block 530, generate, by the third module, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view. In an embodiment, the occlusion-aware parametric road layout prediction includes a prediction boundary such as, for example, but not limited to, a road boundary, a crosswalk boundary, a sidewalk boundary, a lane boundary, and an object (person, other vehicle, building, etc.) boundary that can be collided with, wherein one or more members of the group are originally occluded by a respective foreground element in the perspective view.

[0045] At block 540, control a vehicle system of a vehicle for accident avoidance responsive to the occlusion-aware parametric road layout prediction. The vehicle system can be any one or more of a braking system, an accelerating system, a stability system, a steering system, and so forth.

[0046] A description will now be given of the framework of the present invention, in accordance with an embodiment of the present invention.

[0047] The model includes three modules. The first Perspective Semantics (PS) module 310 inputs the RGB image and outputs the Occlusion-reasoned pixel-level Semantics in Perspective view (OSP). The second Top-view Semantics (TS) module 330 projects OSP into top-view and learns to hallucinate/complete pixel-level top-view semantics on out-of-view as well as noisy regions, which is referred to as Hallucinated Semantics in Top-view (HST). The last Top-view Parametric Prediction module 340 parses the HST and provides predictions on road layout related attributes in top-view.

[0048] A description will now be given regarding a full model, in accordance with an embodiment of the present invention.

[0049] Here, the present invention focuses on model structure and assumes that supervisions are available for each module. It is highlighted that the only annotations required are in parametric form. Though it is assumed that dense pixel-level supervision/annotations are available, they are actually from the generation process of the present invention rather than human annotations. Let it first be assumed that there is a data set $D^r = \{I, x^p, x, \Theta\}_{i=1}^{N^r}$ of N^r samples. Specifically, $I \in \mathbb{R}^{H \times W \times 3}$ is the RGB perspective image, where H and W are image height and width. $x^p \in \mathbb{R}^{H \times W \times (C+1)}$ denotes semantic segmentation map in perspective view, with $(C + 1)$ semantic categories (“road”, “sidewalk”, “lane boundaries”, “crosswalks” and “foreground”). C denotes the number of background categories and equals to 4 in experiments, although other numbers can be used. The semantic top-views are denoted as $x \in \mathbb{R}^{h \times w \times (C+1)}$, with spatial dimensions $h \times w$, including the same $C + 1$ semantic categories as well. Finally, for each data sample, its corresponding scene attributes are denoted as Θ .

[0050] The full model is defined as:

$$\Theta = f^{\text{full}}(I) = (f \circ f^{ts} \circ f^{ps})(I),$$

where $\circ$ defines a function composition. f^{ps} , f^{ts} and f correspond to three modules.

[0051] A description will now be given regarding the perspective semantics module, in accordance with an embodiment of the present invention.

[0052] The goal of Perspective Semantics Module 310 is to provide per-pixel occlusion-reasoned semantics in perspective view, or OSP, on the given RGB. Note that compared to traditional semantic segmentation models that predict semantics on visible pixels only, the Perspective Semantics module 310 focuses on predicting both

visible and occluded background classes. This requires the module to learn to predict semantics with contextual cues rather than fully relying on visible information.

[0053] Assuming that such desired ground-truths, x^p , are available in a dataset, the structure of HRNetV2-W18 is followed as a semantic segmentation backbone as it achieves very good trade-offs between accuracy and efficiency. Formally, given an $I \in \mathbb{R}^{H \times W \times 3}$, the PS module would output a $x^p \in \mathbb{R}^{H \times W \times (C+1)}$, which denotes the probability of each pixel belonging to a specific category.

[0054] The PS module 310 is defined as follows:

$$x^p = f^{PS}(I)$$

[0055] A description will now be given regarding the top-view semantics module 210, in accordance with an embodiment of the present invention.

[0056] The second module, i.e., the top-view Semantics module 260, inputs the OSP and learns to explicitly project the semantics in perspective view to top-view. Formally, this module takes the input $x^p \in \mathbb{R}^{H \times W \times (C+1)}$ and outputs $x \in \mathbb{R}^{h \times w \times (C+1)}$. Again, x denotes the probabilistic map in top-view, where h and w are the height and width of top-view map. Given camera parameters, it seems that the projection is trivial if one has the depth estimation available, say, via a depth network. However, there are mainly three problems in this simple solution. First of all, pixel-wise dense depth estimation with single image as input can be inaccurate. With or without ground-truth supervision, depth accuracy can be unsatisfactory on far away regions and boundaries. Furthermore, standard single image depth networks typically do not reason depth in occluded regions, which is nevertheless required for occlusion-aware projection in accordance with the present invention. Second, resolution is low for distant regions and thus may lead to sparser/noisier semantics in top-view. Lastly, top-view semantics on close-by regions can be incomplete due to limited field of view.

[0057] In view of these challenges, the present invention first makes use of the prior that the road forms nearly a plane, which facilitates the projection without requiring depth estimation. This step can be considered to be a transformation module. Second, the present invention provides a hallucination module 330 to address the sparser/noisy semantics and incomplete predictions on far-away and close-by regions in top-view.

[0058] Specifically, known camera intrinsics and extrinsics with respect to the ground plane are assumed; this is a mild assumption since they could be obtained a prior via calibration in advance. As such, it is well-known that one can compute the homography that maps each pixel in the perspective view to the BEV view and vice-versa. This yields the transformation module.

[0059] After the transformation module 320 that maps the OSP to top-view, the hallucination module 330 then learns to predict the unseen far away regions as well as recover the noisy semantics with contextual information in top-view. A shallow version, e.g. 5-layer encoder and decoder, is used as the hallucination module 330. Note that the input and output of hallucination module 330 is of the same size, or $h \times w \times (C + 1)$.

[0060] In summary, the TS module is defined as follows:

$$x = f^{\text{ts}}(x^p) = (f^{\text{halln}} \circ f^{\text{trans}})(x^p)$$

where f^{halln} and f^{trans} are the transformation module 320 and hallucination module 330, respectively.

[0061] A description will now be given regarding the top view parametric prediction module 340, in accordance with an embodiment of the present invention.

[0062] Given the Hallucinated Semantics in Top-view (HST), the next step is to predict the layout attributes. In brief, the Top-view Parametric Prediction (TPP) module 340 maps the HST x into the scene model parameters Θ . The Θ includes three

groups, or Θ_b for 14 binary, Θ_m for 2 multi-class and Θ_c for 10 continuous attributes of scene model, respectively. Binary attributes include information like whether the road is one-way or not. The number of lanes on the left hand-side of the ego car is an example of the multi-class attribute and distance to right side-road can be one of the continuous attributes.

[0063] The TTP module 340 is defined as follows:

$$\Theta = f(x) = (h \circ g)(x),$$

where h and g are multi-layer perceptron (MLP) and convolutional neural networks respectively. Specifically, h is implemented as a multi-task network with three separate predictions n_b , n_m , and n_c for each of the parameter groups Θ_b , Θ_m and Θ_c of the scene model.

[0064] The TTP module 340 is also able to exploit the simulate data. One can easily extend the current design to a hybrid version that leverages the rich and large set of annotations from simulated data during training.

[0065] A description will now be given regarding model training, in accordance with an embodiment of the present invention.

[0066] It is assumed that supervisions are all available for each module. A description will be given regarding how to collect such dataset when having the accessibility to human annotations for parametric layout Θ only. Then, an introduction is made into how to exploit deep supervision during training.

[0067] The full loss function L is defined as:

$$\mathcal{L} = \lambda \times L^{\text{tp}} + \gamma \times L^{\text{ts}} + \beta \times L^{\text{ps}}$$

where λ , γ and β are the weights for each module.

[0068] A description will now be given regarding loss functions on the top-view Parametric Prediction Module, in accordance with an embodiment of the present

invention. Since Θ and I are already available, the loss function of parametric prediction is defined as follows:

$$\mathcal{L}^{\text{tp}} = \sum_{i=1}^{N^r} \text{BCE}(\Theta_{b,i}^r, n_{b,i}^r) + \text{CE}(\Theta_{m,i}^r, n_{m,i}^r) + \ell_1(\Theta_{c,i}^r, n_{c,i}^r),$$

where (B)CE is the (binary) cross-entropy loss and $\{\Theta, n\}, i$ denotes the i -th sample in the data set. For regression, continuous variables are discretized into 100 bins by convolving a Dirac delta function centered at Θ_c with a Gaussian of fixed variance. With the help of multi-modal predictions, the model can be easily extended with graphical models.

[0069] A description will now be given regarding a top view semantics module, in accordance with an embodiment of the present invention.

[0070] Unlike the straight-forward design in parametric space, the second module requires per-pixel supervision in top-view. To this end, it is proposed to exploit a rendering function that generates pixel-wise semantics from parametric annotations. Specifically, for each Θ , a x map is rendered.

[0071] A description will now be given regarding loss functions on the top view semantics module.

[0072] The loss functions for TS module is defined as follows:

$$\mathcal{L}^{\text{ts}} = \sum_{i=1}^{N^r} \text{CE}(x_i, \hat{x}_i)$$

where $\hat{x}_i$ and x_i denotes the predictions and the rendered ground-truth of the top-view semantics of i -th sample in D^r .

[0073] A description will now be given regarding the Perspective Semantics Module, in accordance with an embodiment of the present invention.

[0074] As long as the top-view semantics $\mathcal{x}$ are obtained, they can be mapped back to perspective view with camera parameters as well as plane assumption. Again, this is achieved by computing the homography between the perspective and BEV views.

[0075] A back projection can indeed recover the occluded regions and aligns the semantics well with respect to the input image in perspective view.

[0076] A description will now be given regarding loss functions on the Perspective Semantics Module, in accordance with an embodiment of the present invention.

[0077] Similarly, the loss function for PS module is as follows:

$$\mathcal{L}^{\text{PS}} = \sum_{i=1}^{N^{\text{r}}} \text{CE}(x_i^{\text{p}}, \hat{x}_i^{\text{p}})$$

where $\hat{x}_i^{\text{p}}$ and x_i^{p} denotes predictions in accordance with the present invention and the back-projected ground-truth of the perspective semantics of i -th sample in D^{r} .

[0078] The present invention may be a system, a method, and/or a computer program product at any possible technical detail level of integration. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.

[0079] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable

programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals *per se*, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0080] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0081] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as SMALLTALK, C++ or the like, and conventional

procedural programming languages, such as the “C” programming language or similar programming languages. The computer readable program instructions may execute entirely on the user’s computer, partly on the user’s computer, as a stand-alone software package, partly on the user’s computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user’s computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention.

[0082] Aspects of the present invention are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[0083] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified

in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[0084] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0085] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the

block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[0086] Reference in the specification to “one embodiment” or “an embodiment” of the present invention, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well any other variations, appearing in various places throughout the specification are not necessarily all referring to the same embodiment.

[0087] It is to be appreciated that the use of any of the following “/”, “and/or”, and “at least one of”, for example, in the cases of “A/B”, “A and/or B” and “at least one of A and B”, is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of both options (A and B). As a further example, in the cases of “A, B, and/or C” and “at least one of A, B, and C”, such phrasing is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of the third listed option (C) only, or the selection of the first and the second listed options (A and B) only, or the selection of the first and third listed options (A and C) only, or the selection of the second and third listed options (B and C) only, or the selection of all three options (A and B and C). This may be extended, as readily apparent by one of ordinary skill in this and related arts, for as many items listed.

[0088] The foregoing is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the invention disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted

according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the present invention and that those skilled in the art may implement various modifications without departing from the scope and spirit of the invention. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the invention. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A computer-implemented method for road layout prediction, comprising:
segmenting (510), by a first processor-based element, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues;
learning (520), by a second processor-based element, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption; and
generating (530), by a third processor-based element, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view
2. The computer-implemented method of claim 1, further comprising injecting a respective loss into each of the segmenting, learning and generating steps, the respective loss being minimized in each of the steps to generate the occlusion-aware parametric road layout prediction.
3. The computer-implemented method of claim 1, wherein said generating step comprises receiving pixel-level human annotations only for the road layout related attributes in the top view, and wherein said segmenting and learning steps are performed without human annotations.
4. The computer-implemented method of claim 1, wherein said segmenting and learning steps use pixel-level supervision signals.

5. The computer-implemented method of claim 1, wherein the occlusion-aware parametric road layout prediction comprises a prediction boundary selected from a group consisting of a road boundary, a crosswalk boundary, a sidewalk boundary, and a lane boundary, wherein one or more members of the group are originally occluded by a respective foreground element in the perspective view.

6. The computer-implemented method of claim 1, wherein the pixel-level semantic segmentation results and the top view of the RGB image are occlusion-reasoned to indicate items occluded in the perspective view of the RGB image.

7. The computer-implemented method of claim 1, further comprising controlling (540) a vehicle system of a vehicle for accident avoidance responsive to the occlusion-aware parametric road layout prediction.

8. The computer-implemented method of claim 7, wherein the vehicle system is selected from a group consisting of a braking system, an accelerating system, a stability system, and a steering system.

9. The computer-implemented method of claim 1, wherein the road plane assumption assumes the road is represented by a plane in three-dimensional space that, in turn, is represented by its normal and distance to a camera.

10. The computer-implemented method of claim 1, wherein the method is performed by an Advanced Driver Assistance System.

11. A computer program product for road layout prediction, the computer program product comprising a non-transitory computer readable storage medium having program instructions embodied therewith, the program instructions executable by a computer to cause the computer to perform a method comprising:

segmenting (510), by a first processor-based element of the computer, an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues;

learning (520), by a second processor-based element of the computer, a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption; and

generating (530), by a third processor-based element of the computer, an occlusion-aware parametric road layout prediction for road layout related attributes in the top view

12. The computer program product of claim 11, wherein the method further comprises injecting a respective loss into each of the segmenting, learning and generating steps, the respective loss being minimized in each of the steps to generate the occlusion-aware parametric road layout prediction.

13. The computer program product of claim 11, wherein said generating step comprises receiving pixel-level human annotations only for the road layout related attributes in the top view, and wherein said segmenting and learning steps are performed without human annotations.

14. The computer program product of claim 11, wherein said segmenting and learning steps use pixel-level supervision signals.

15. The computer program product of claim 11, wherein the occlusion-aware parametric road layout prediction comprises a prediction boundary selected from a group consisting of a road boundary, a crosswalk boundary, a sidewalk boundary, and a lane boundary, wherein one or more members of the group are originally occluded by a respective foreground element in the perspective view.

16. The computer program product of claim 11, wherein the pixel-level semantic segmentation results and the top view of the RGB image are occlusion-reasoned to indicate items occluded in the perspective view of the RGB image.

17. The computer program product of claim 11, wherein the method further comprises controlling (540) a vehicle system of a vehicle for accident avoidance responsive to the occlusion-aware parametric road layout prediction.

18. The computer program product of claim 17, wherein the vehicle system is selected from a group consisting of a braking system, an accelerating system, a stability system, and a steering system.

19. The computer program product of claim 11, wherein the road plane assumption assumes the road is represented by a plane in three-dimensional space that, in turn, is represented by its normal and distance to a camera.

20. A computer processing system for road layout prediction, comprising:

a memory device (140) for storing program code; and

a processor device (110) for running the program code to:

segment an RGB image to output pixel-level semantic segmentation results for the RGB image in a perspective view for both visible and occluded pixels in the perspective view based on contextual clues;

learn a mapping from the pixel-level semantic segmentation results for the RGB image in the perspective view to a top view of the RGB image using a road plane assumption; and

generate an occlusion-aware parametric road layout prediction for road layout related attributes in the top view.

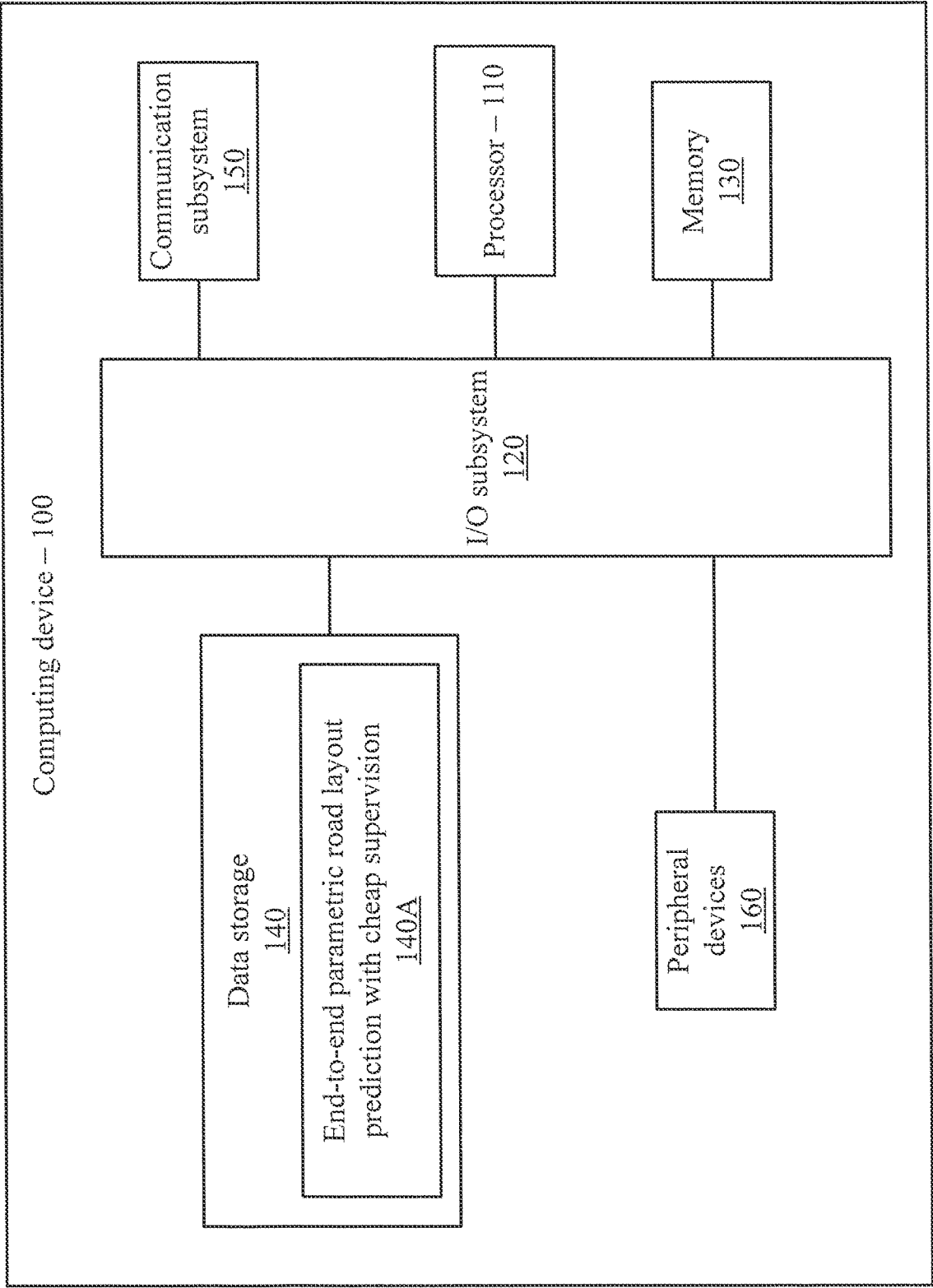


FIG. 1

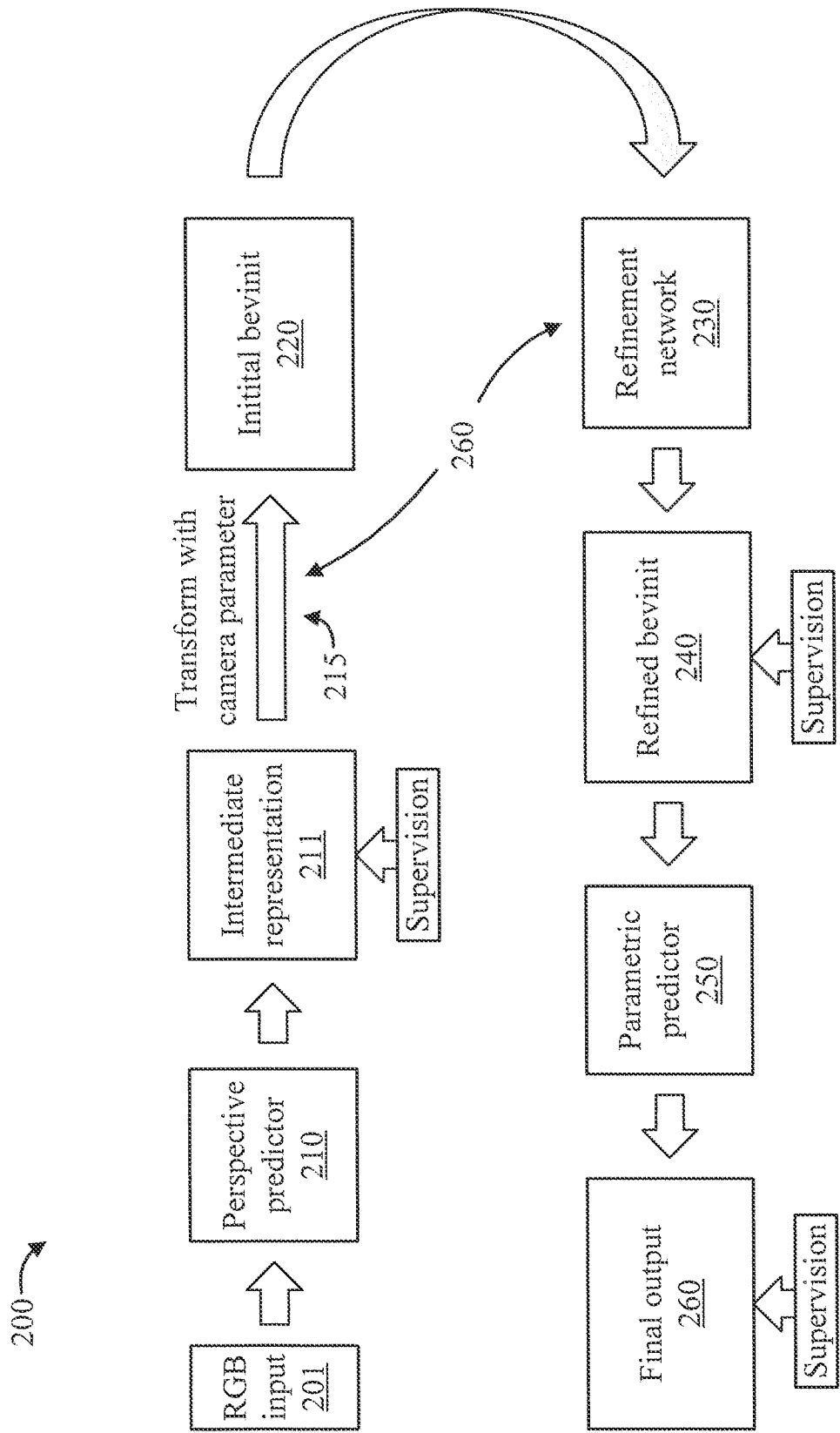


FIG. 2

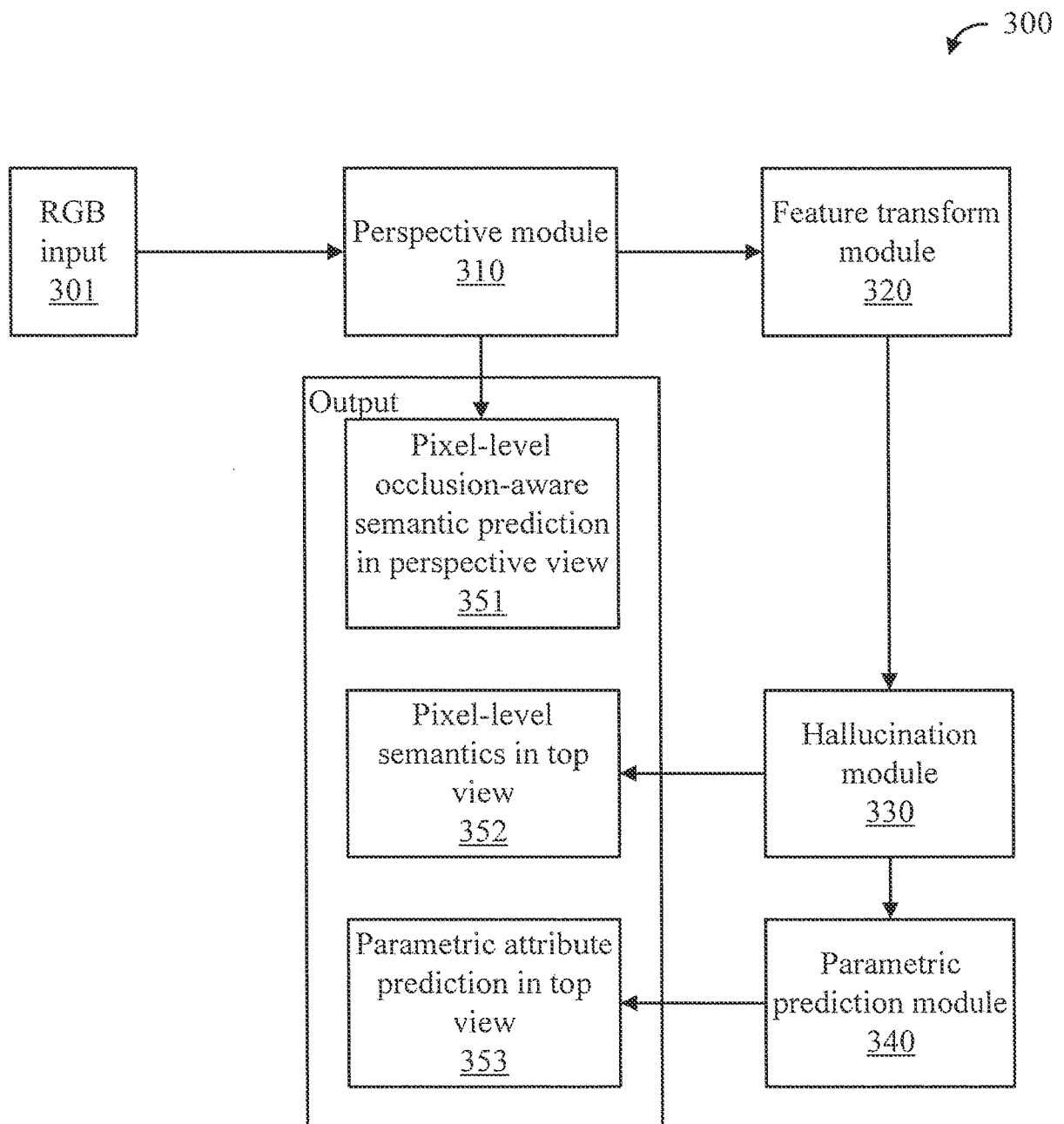


FIG. 3

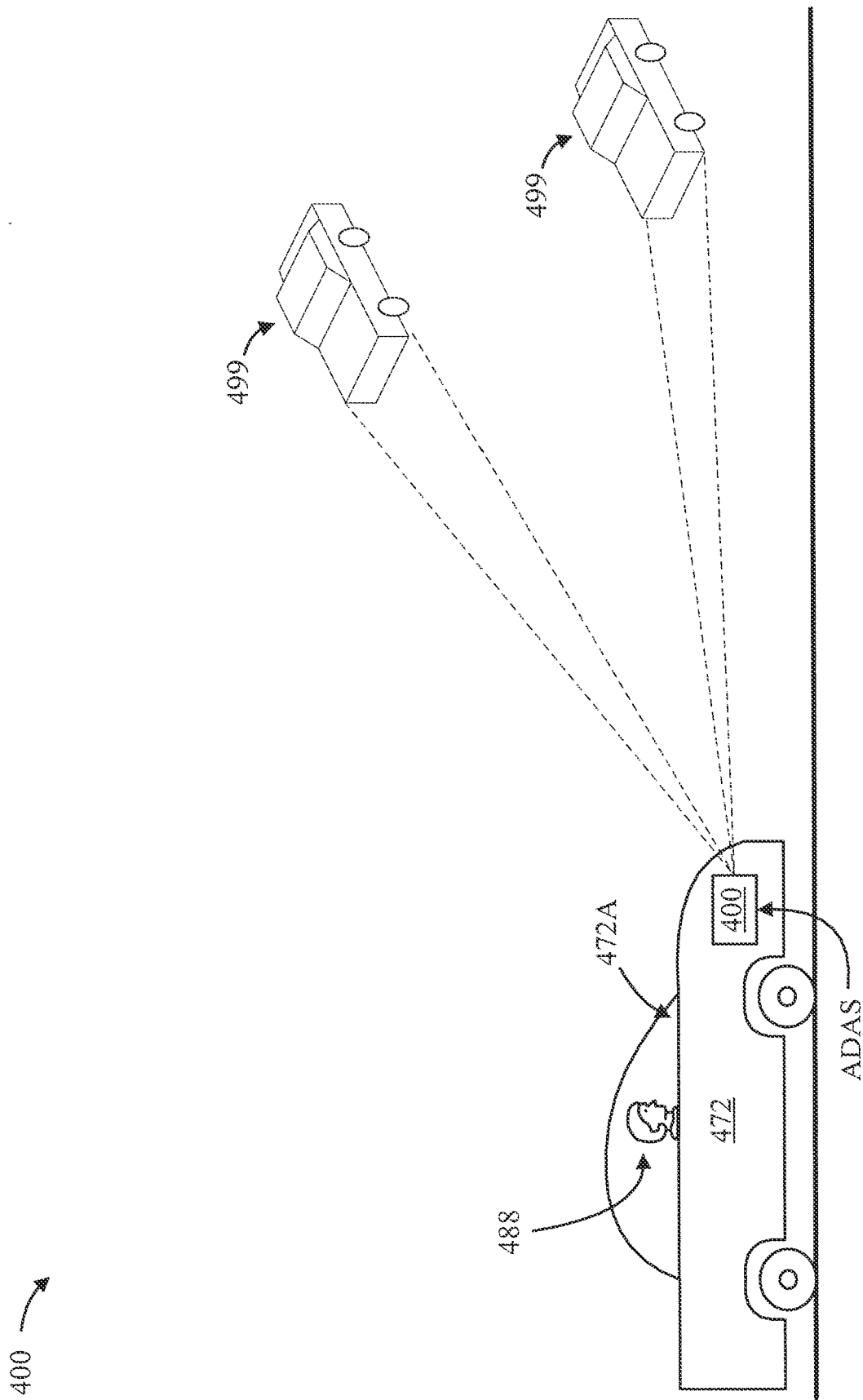


FIG. 4

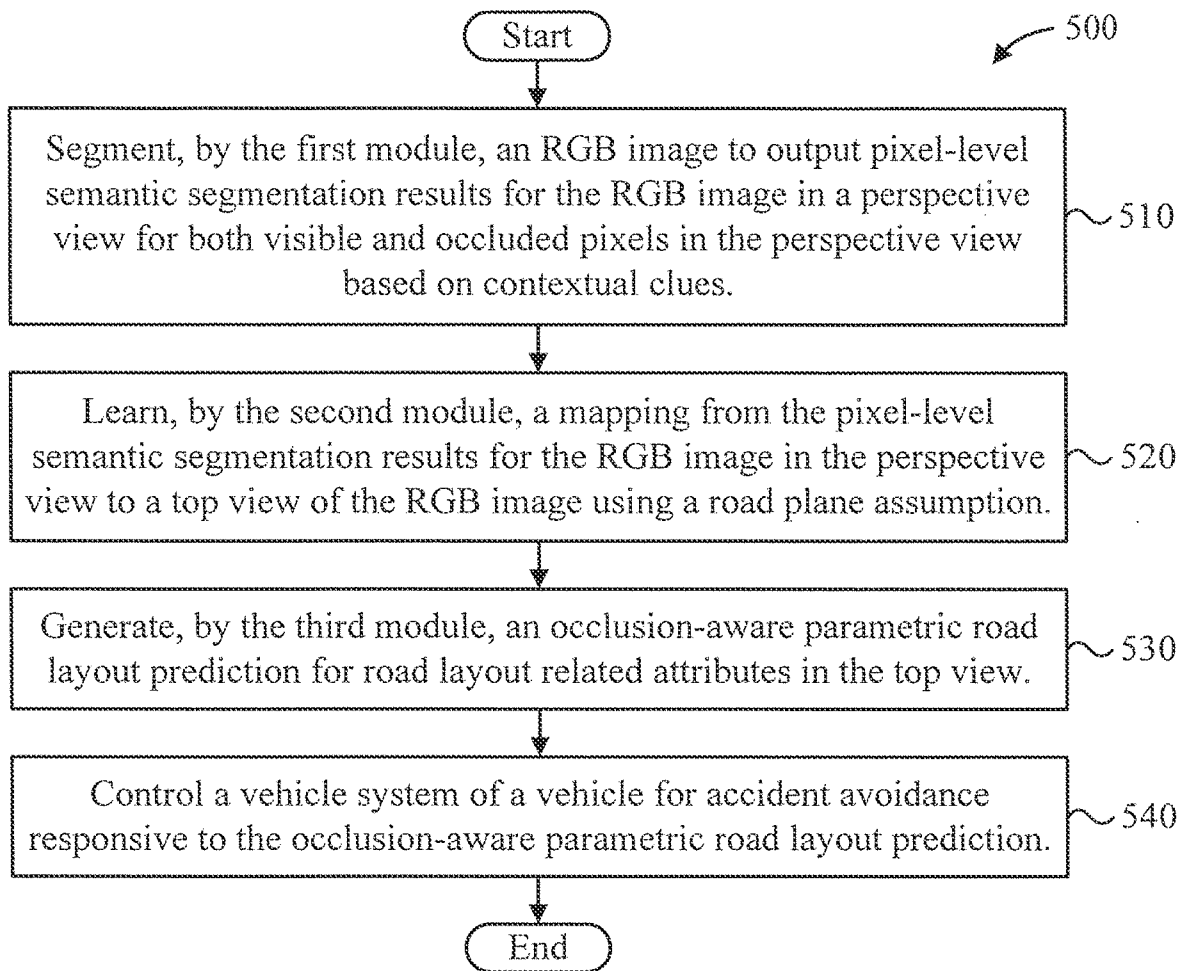


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2021/058615

A. CLASSIFICATION OF SUBJECT MATTER**G06K 9/00**(2006.01)i; **G06N 20/00**(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K 9/00(2006.01); G05D 1/02(2006.01); G06F 17/30(2006.01); G06K 9/46(2006.01); G06K 9/62(2006.01);
G06T 17/05(2011.01); G06T 7/73(2017.01); G08G 1/16(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: road, prediction, segmenting, mapping, top view, occlusion

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2020-0050900 A1 (NEC LABORATORIES AMERICA, INC.) 13 February 2020 (2020-02-13) paragraphs [0016]-[0056], claims 1-2 and figures 1-6	1-20
A	US 2019-0095722 A1 (SAMSUNG ELECTRONICS CO., LTD.) 28 March 2019 (2019-03-28) claims 1, 14-15 and figures 1-9	1-20
A	US 2016-0004916 A1 (WENMING ZHENG et al.) 07 January 2016 (2016-01-07) claims 1-2 and figures 1-7	1-20
A	US 2018-0232946 A1 (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 16 August 2018 (2018-08-16) claims 1-3 and figures 1-10	1-20
A	US 2020-0160069 A1 (NAUTO, INC.) 21 May 2020 (2020-05-21) claims 1-2 and figures 1-7	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

07 March 2022

Date of mailing of the international search report

07 March 2022

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsu-ro, Seo-gu, Daejeon
35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

PARK, Hye Lyun

Telephone No. +82-42-481-3463

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2021/058615

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2020-0050900	A1	13 February 2020	JP	2021-528798	A	21 October 2021
				WO	2020-050934	A2	12 March 2020
				WO	2020-050934	A3	18 June 2020
US	2019-0095722	A1	28 March 2019	CN	109584578	A	05 April 2019
				EP	3462377	A2	03 April 2019
				EP	3462377	A3	08 May 2019
				JP	2019-067364	A	25 April 2019
				KR	10-2019-0036844	A	05 April 2019
				US	10740626	B2	11 August 2020
US	2016-0004916	A1	07 January 2016	CN	104937608	A	23 September 2015
				CN	104937608	B	07 December 2018
				EP	2962249	A1	06 January 2016
				US	10936883	B2	02 March 2021
				WO	2014-131193	A1	04 September 2014
US	2018-0232946	A1	16 August 2018	CN	105069842	A	18 November 2015
				EP	3319048	A1	09 May 2018
				JP	2018-509668	A	05 April 2018
				JP	6568587	B2	28 August 2019
				KR	10-1932623	B1	27 December 2018
				KR	10-2017-0103763	A	13 September 2017
				US	10643378	B2	05 May 2020
				WO	2017-020465	A1	09 February 2017
US	2020-0160069	A1	21 May 2020	EP	3654241	A1	20 May 2020
				JP	2020-115322	A	30 July 2020