

US Patent & Trademark Office

Patent Public Search | Text View

United States Patent Application Publication

20250265833

Kind Code

A1

Publication Date

August 21, 2025

Inventor(s)

Li; Yumeng et al.

TECHNIQUE FOR EXTRACTING FEATURES OF AN ENVIRONMENT FROM IMAGE DATA

Abstract

A technique for extracting features of an environment from image data is provided. A computer implemented method includes receiving data indicative of a visual domain of an environment; and generating a visual domain textual prompt based on the received data indicative of the visual domain of the environment. The method further includes receiving image data representative of the environment. The method further includes extracting, in particular local, features of the environment from the received image data. The extracting of the, in particular local, features is performed by a conditional feature extracting model. The extracting of the, in particular local, features is conditioned by the generated visual domain textual prompt.

Inventors: Li; Yumeng (Barcelona, ES), Khoreva; Anna (Berlin, DE), Zhang; Dan (Leonberg, DE)

Applicant: Robert Bosch GmbH (Stuttgart, DE)

Family ID: 1000008450014

Appl. No.: 19/051242

Filed: February 12, 2025

Foreign Application Priority Data

EP 24 15 7953.1

Feb. 15, 2024

Publication Classification

Int. Cl.: G06V10/82 (20220101); G06F40/40 (20200101); G06V10/764 (20220101); G06V10/772 (20220101)

U.S. Cl.:

Background/Summary

CROSS REFERENCE

[0001] The present application claims the benefit under 35 U.S.C. § 119 of European Patent Application No. EP 24 25 7953.1 filed on Feb. 15, 2024, which is expressly incorporated herein by reference in its entirety.

FIELD

[0002] A technique for extracting features of an environment from image data and a technique for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data are provided according to the present invention. The techniques of the present invention comprise methods, computing devices, a system, a computer program product, and a computer readable storage medium.

BACKGROUND INFORMATION

[0003] Recent large-scale text-to-image generative models, e.g., Stable Diffusion (SD, as described by R. Rombach et al. in “High-Resolution Image Synthesis with Latent Diffusion Models”, arXiv:2112.10752v2 [cs.CV], which is incorporated herein by reference), trained on a large amount of text-image pairs, have demonstrated an impressive performance and ability to generate diverse and creative images given a text prompt. In particular, SD is the state-of-the-art open-source vision-language generative model trained over billions of text-image pairs. SD is a latent diffusion model, which is a special case of the diffusion model trained in a certain latent space instead of the original image space. More specifically, following P. Esser et al., “Taming Transformers for High-Resolution Image Synthesis”, arXiv:2012.09841v3 [cs.CV] (known for short also as VQ-GAN), which is incorporated herein by reference, and as schematically illustrated in FIG. 5, SD firstly trains an autoencoder, where the image is encoded into the latent space Z, and decoded back to reconstruct the given image. In a second stage, a diffusion model is trained in this latent space Z. Due to being trained on a large amount of data, SD has a learned rich representation.

[0004] Recent work by W. Zhao et al. in “Unleashing Text-to-Image Diffusion Models for Visual Perception (VPD)”, arXiv:2303.02153v1 [cs.CV], which is incorporated herein by reference, is built on top of SD and exploits the pretrained knowledge of a denoising UNet to provide semantic guidance in various visual perception tasks, e.g., semantic segmentation, and depth estimation. W. Zaho et al. explored to fine-tune SD and simultaneously train a task-specific decoder for various visual perception tasks (e.g., semantic segmentation, depth estimation), as schematically shown in FIG. 6.

[0005] However, SD is finetuned by W. Zaho et al. on a specific training set and only evaluated in the same trained domain at inference time.

SUMMARY

[0006] In the following, the techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data are disclosed with respect to methods of the present invention as well as with respect to computing devices of the present invention. Features, advantages, or alternative embodiments herein can be assigned to the other objects of the present invention (e.g., the system comprising a computing device, the computer program or a computer program product), and vice versa. In other words, elements of the computing devices can be improved with features disclosed in the context of the methods of the present invention. In this case, the functional features of the method of the present invention are embodied by structural units of the system of the present

invention and vice versa, respectively.

[0007] As to a first method aspect of the present invention, a computer-implemented method for extracting features of an environment from image data is provided. According to an example embodiment of the present invention, the method comprises a step of receiving data indicative of a visual domain of an environment. The method further comprises a step of generating a visual domain textual prompt based on the received data indicative of the visual domain of the environment. The method further comprises a step of receiving image data representative of the environment. The method still further comprises a step of extracting (in particular local) features of the environment from the received image data. The extracting of the (in particular local) features is performed by a conditional feature extracting model and is conditioned by the generated visual domain textual prompt.

[0008] The method of the present invention may be used for providing a model for image generation, which may adapt itself flexibly based on the visual domain textual prompt, indicative of the visual domain of the environment.

[0009] By the computer-implemented technique of the present invention, performing feature extraction of an environment from the received image data, and preferably performing a visual perception task based on the extracted features, which is more robust and generalizable to various visual domains (e.g., different weather condition, lighting conditions, and/or new countries) is enabled. Alternatively or in addition, an overfitting to training datasets (e.g., comprising and/or based on the extracted features conditioned by the visual domain textual prompt, in particular as ground truth) for training a model (e.g., a neural network, NN, and/or an artificial intelligence, AI) for performing the feature extraction and/or the visual perception task may be avoided. E.g., an accurate result of the feature extraction and/or of the visual perception task (in particular object detection, depth estimation, classification, and/or semantic segmentation) can be critical for safe automated driving (AD), and/or for safely operating a robot (e.g., in a manufacturing environment, and/or in a home automation).

[0010] An AD driving functionality may comprise a full automation (also denoted as autonomous driving, in particular L5), an at least partial (e.g., high, L4, conditional, L3, or partial, L2) automation, and/or an assisted driving automation (e.g., L1, e.g., comprising adaptive cruise control, ACC).

[0011] Any textual prompt may also be denoted as text prompt and/or language prompt. Alternatively or in addition a textual prompt may be provided and/or processed in a text-based and/or acoustic-based format.

[0012] The visual domain textual prompt (also: domain indicator) can improve on the (in particular visual) domain flexibility of the feature extraction and/or the visual perception task (e.g., by using knowledge on an expected underexposure or overexposure).

[0013] A (e.g., pretrained) model for performing the feature extraction and/or the visual perception task, which is trained on training datasets (in particular comprising the result of the feature extraction conditioned by the visual domain textual prompt, and/or the result of the visual perception task based on the feature extraction), can be significantly improved, in particular by enabling a transfer of (in particular rich) prior knowledge and/or fast training convergence.

[0014] The environment may also be denoted as (in particular physical, real and/or natural) scene, setting, and/or ambience. The environment represented by the received image data may comprise an environment of a (e.g., AD) vehicle, of a robot, of a surveillance system, and/or of a person. In particular, the environment may represent a traffic scene. Alternatively or in addition, the environment may be real, natural, rural, built, urban, and/or spatial. E.g., the environment of the (e.g., AD) vehicle may comprise a road layout, the environment of the robot may comprise a manufacturing site, and the environment of the surveillance system, and/or of the person, may comprise a building structure. E.g., for AD applications, the environment may comprise an outside environment, in particular a street view. Further exemplarily, for a robot, the environment may

comprise a factory site having an inside and/or outside area.

[0015] The image data representative of the environment may be received as digital (and/or analog) data, in particular from a (e.g., video) camera. Alternatively or in addition, the image data of the environment may be acquired by means of a (e.g., video and/or thermal) camera, radar, LiDAR, ultrasonic sensors, and/or motion sensors.

[0016] The visual domain may also be denoted as visibility conditions, optical conditions, and/or may comprise an indication of a visual (also: optical) setting of the environment (e.g., in view of lighting and/or weather conditions. Alternatively, or in addition the visual domain may be generated from a geographical identifier, and/or may be, in particular country-specific, such as lighting conditions based on the time of acquiring the image data relative to local sunrise and sunset times, layout of traffic signs and/or streets for outside environments, and/or types of natural and/or artificial lighting for indoors environments). The lighting conditions (briefly: lighting) may alternatively be denoted as illumination and/or brightness.

[0017] The data indicative of the visual domain may comprise digital data, e.g., a result of a brightness determination of imaging data. Alternatively or in addition, the digital data may be converted from a (e.g., manual) input (e.g., a selected brightness and/or weather condition), and/or may be compressed, e.g., in terms of a (in particular manually and/or automatically) selected one of a number of predetermined visual domains (e.g., arranged in a table).

[0018] According to an example embodiment of the present invention, the step of generating the visual domain textual prompt may be performed, at least partly, by a contrastive language-image pretraining (CLIP) encoder. In some embodiments, a pretrained Visual Question Answering (VQA) model, e.g., mPLUG, may first receive the data indicative of the visual domain of the environment and provide a result, in particular as a probability of a visual domain class out of a discrete set of visual domain classes. In one example embodiment of the present invention, the visual domain class with the highest probability may be selected as the visual domain, for which the visual domain textual prompt is generated. In an alternative embodiment, the visual domain, for which the visual domain textual prompt is generated, may correspond to a weighted combination of a number K (e.g., with $K=2$ corresponding to the two highest probabilities) of visual domain classes with the K highest probabilities.

[0019] The conditional feature extracting model may comprise a generative model, and/or a (in particular large-scale) text-to-image generative artificial intelligence (AI).

[0020] The extracted features may represent the environment locally. Alternatively or in addition, the extracted features may comprise objects, shapes and/or colors.

[0021] Since the feature extraction process is conditioned on the textual prompt indicative of the visual domain, e.g., location and/or country information may be relevant for the extracted feature. E.g., some local traffic signs may be different from location to location, and/or driving on the left-hand side or on right hand side may be used as prior knowledge for the model prediction. The feature extraction may, e.g., additionally be conditioned on an environmental class textual prompt, wherein the environmental class relates to, e.g., the country (e.g., for a typical road layout, and/or positioning and/or appearance of a traffic sign and/or of a traffic light), a rural environment, and/or an urban environment (e.g., in view of typical cultivation and/or buildings).

[0022] The (in particular locally) extracted features may be supplemented by cross-attention maps. The cross-attention maps may relate different patches of the (in particular locally) extracted features. Alternatively or in addition, the cross-attention maps can facilitate consistency across the environment comprised in the image data, in particular as a result of a subsequently performed visual perception task.

[0023] The extracted features of the environment, optionally supplemented with the cross-attention maps, may be provided for output.

[0024] According to an example embodiment of the present invention, the method may further comprise a step of performing a visual perception task on the received image data, representing the

environment, based on the extracted (in particular local) features. The result of the performed visual perception task may comprise an (e.g., weighted) average over features, in particular at different timesteps. Alternatively or in addition, a result of the performed visual perception task may be provided and/or output.

[0025] The visual perception task may be performed by a task-specific decoder.

[0026] By performing the visual perception task, the environment captured in the image data can be analyzed in a sufficiently accurate manner to enable, e.g., AD, the use of an autonomous robot, and/or access control.

[0027] According to an example embodiment of the present invention, the visual perception task may comprise a depth estimation, an object detection, a classification, and/or a semantic segmentation. So, the method may be used for image classification. Image classification may comprise object detection and segmentation, in particular semantic segmentation. The classification may be based on low level features of the image data, like, e.g. pixels, edges and/or corners.

[0028] By the depth estimation, an estimate of a distance (e.g., to an obstacle for an, in particular AD, vehicle, and/or a, in particular autonomous, robot) can be performed. Alternatively or in addition, by the object detection and/or by the classification, the detection of a path, a road surface, a traffic sign, a pedestrian, a vehicle, an (e.g., further) obstacle (e.g., a tree), the sky, and/or a destination is enabled. Further alternatively or in addition, performing the classification allows to recognize, e.g., a traffic sign and/or an obstacle, and plan a path (and/or trajectory) accordingly. Further alternatively or in addition, by the semantic segmentation, the (e.g., full) semantic content of the environment can be provided and the path (and/or trajectory) planning of the (e.g., AD) vehicle and/or the (e.g., autonomous) robot can be improved in accuracy, in particular depending on a real-time situation of the environment.

[0029] Alternatively or in addition, according to an example embodiment of the present invention, the visual perception task may comprise, e.g., performing a regression analysis to determine a (in particular continuous) value, such as a distance (e.g., according to the depth estimation), a speed (also: velocity), an acceleration, and/or tracking an object and/or a person, in particular by taking into account edges and/or pixel attributes for image data, and/or by analyzing (in particular image) data, e.g., scalar time series, for performing the corresponding measurements (e.g., of the speed and/or acceleration) and/or for the tracking.

[0030] By performing the visual perception task, an anomaly detection, and optionally controlling, e.g., a (in particular AD) vehicle and/or a (in particular autonomous) robot after the anomaly detection in a fail-safe mode can be provided.

[0031] E.g., in an example embodiment of the present invention, which is applied to AD, some anomaly may show up on the road, such as described by Dan Zhang et al. in “Anomaly-Aware Semantic Segmentation via Style-Aligned OoD Augmentation”, arXiv:2308.09965v1 [cs.CV], which is incorporated herein by reference. E.g., in case an accident happened, “potential debris on the road” may be indicated in the (e.g., environmental class) textual prompt.

[0032] Performing the visual perception task on the image data may be conditioned on the generated visual domain textual prompt.

[0033] According to some embodiments of the present invention, the extracting of the (in particular local) features may be performed for multiple timesteps, in particular in relation to generating latent image data with different noise levels (briefly: noisy latents). The visual perception task may be performed based on the combined (and/or aggregated, and/or weighted) extracted features from the multiple timesteps.

[0034] Alternatively or in addition, according to an example embodiment of the present invention, the conditioning on the generated visual domain textual prompt may comprise a conditioning on a combined (and/or aggregated, and/or weighted) visual domain textual prompt associated with multiple timesteps, in particular in relation to generating the latent image data.

[0035] The received data indicative of the visual domain may comprise environmental sensor data,

position data determined by means of a positioning system (in particular by a satellite navigation system), manually input data, and/or electronically available information data.

[0036] The environmental sensor data may be acquired by means of an environmental sensor. The environmental sensor may comprise an optical sensor, an acoustic sensor, a rain sensor, a humidity sensor, a thermal sensor, a temperature sensor, a wind sensor, a radar sensor, a LiDAR sensor, an ultrasonic sensor and/or a motion sensor. Alternatively or in addition, the environmental sensor may be mounted on the (e.g., AD) vehicle, the (e.g., autonomous) robot, and/or may be stationary within the environment. By means of the environmental sensor, information on the visual domain such as a weather condition (e.g., sunny, rainy and/or foggy, in particular using the humidity sensor) and/or a lighting condition (e.g., day, night, dawn, and/or twilight) can be advantageously determined in real time locally for (and/or within) the environment.

[0037] The positioning system, in particular the satellite navigation system, may comprise a global navigation satellited system (GNSS), e.g., the global positioning system (GPS) and/or Galileo. Alternatively or in addition, the positioning system may comprise a regional system (e.g., in particular using radio signals), a local positioning system (LPS, e.g., using beacons of radio signals and/or of optical signals, Wi-Fi, and/or Li-Fi), and/or an indoor system (e.g., using ultrasound and/or multilateration, such as Active Bat). By the positioning system, information on the visual domain such as a type of terrain, building development, and/or stationary installations (and thereby, e.g., information as to a lighting set-up such as dimming under a bridge and/or in a tunnel) can be advantageously determined for the environment.

[0038] The manually input data may be received by means of a user interface (UI), in particular a graphical user interface (GUI), and/or by means of a microphone (and/or an acoustic interface) from a user (also: human operator, e.g., a driver of a, in particular AD, vehicle, and/or an operator of a robot). The manually input data may comprise data selected from a drop-down menu and/or any form of discrete selections (e.g., sunny, foggy, rainy, snowy, and/or dark for lighting conditions). Alternatively or in addition, the manually input data may comprise (e.g., natural language) text indicative of the visual domain. The user can, by means of the manually input data, provide information indicative of the visual domain, in particular based on his/her perception and/or based on reading off information from one or more sensors without the need for a (e.g., wireless and/or wired) data connection of the sensors.

[0039] The electronically available information data may be received via a (e.g., wireless and/or wired) interface (in particular via the internet or a cloud-based data storage) and/or may comprise data retrieved from the internet, e.g., a local weather forecast. Thereby, a wide range of, in particular already existing, electronically available information can advantageously be made use of, e.g., for generating the visual domain textual prompt.

[0040] The data indicative of the visual domain may comprise any combination of data types. E.g., the position data may be used to determine the appropriate electronically available information data comprising the local weather forecast for the environment.

[0041] According to an example embodiment of the present invention, the method may comprise a step of receiving data indicative of a class in relation to the environment. The method may further comprise a step of generating an environmental class textual prompt based on the received data indicative of the class in relation to the environment. Extracting (in particular local) features of the environment from the received image data may further be conditioned by the generated environmental class textual prompt.

[0042] Generating the environmental class textual prompt may be performed, at least partially by a CLIP encoder.

[0043] While conventionally, a CLIP encoder is frozen (and/or unchanged), in particular after a training phase, in contrast and according to the computer-implemented technique of the present invention presented herein, generating the environmental class textual prompt, and/or the visual domain textual prompt, may at least partly (e.g., after a first processing of the corresponding

received data by the CLIP encoder) change over time and/or may be performed by a trainable text adapter (e.g., comprising some multi-layer perceptron, MLP, layers).

[0044] By further conditioning the extracting of the (in particular local) features of the environment on the generated environmental class textual prompt, e.g., a classification task, and/or a semantic segmentation task, as visual perception task can be facilitated.

[0045] According to an example embodiment of the present invention, the extracting of features may further be conditioned by the environmental class textual prompt. By the conditioning on the environmental class textual prompt, a classification and/or semantic segmentation as the visual perception task can be facilitated and/or improved.

[0046] According to an example embodiment of the present invention, the step of generating the visual domain textual prompt, and/or the step of generating the environmental class textual prompt may be performed by an adaptive contrastive language-image pretraining (CLIP) encoder. By using an adaptive CLIP encoder for generating the (in particular visual domain and/or environmental class) textual prompt, a generalizability of the computer-implemented technique is enabled and/or improved, in particular in view of a previously untrained (and/or unseen) visual domain and/or a previously untrained (and/or unseen) environmental class.

[0047] According to an example embodiment of the present invention, the conditional feature extracting model may comprise a generative image-to-feature model, which is configured for extracting (in particular local) features of the environment from the received image data of the environment, and a conditioning model, which is configured for encoding the generated visual domain textual prompt for conditioning, and/or controlling, the generative image-to-feature model.

[0048] According to an example embodiment of the present invention, the generative image-to-feature model may be trained (and/or may operate) on a joint vision-language space. The generative image-to-feature model may comprise a vision transformer (ViT, e.g., as described by P. Esser et al. in “Taming Transformers for High-Resolution Image Synthesis”, arXiv:2012.09841v3 [cs.CV]), a (in particular vector graphics, VG) generative adversarial network (GAN, in particular a VG-GAN, e.g., as described by Yong Zheng Ong et al. in “VG-GAN: Conditional GAN Framework for Graphical Design Generation”, 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 781-785, DOI: 10.1109/ICIP46576.2022.9897935, which is incorporated herein by reference), a diffusion model for visual perception (VPD, e.g., as described by Wenliang Zhao et al. in “Unleashing Text-to-Image Diffusion Models for Visual Perception, arXiv:2303.02153 v1 [cs.CV]), and/or Stable Diffusion (SD), in particular a SD denoising UNet.

[0049] According to an example embodiment of the present invention, when performing the method, the generative image-to-feature model may be frozen, and/or need not be further adapted, in the sense that weights of the generative image-to-feature model remain unchanged in a fine-tuning phase and/or an inference phase.

[0050] The optional environmental class textual prompt may condition (e.g., directly) the image-to-feature model (e.g., by feeding the environmental class textual prompt into encoding and/or decoding layers, and/or into optional middle layers).

[0051] According to an example embodiment of the present invention, the generative image-to-feature model may comprise a diffusion model, in particular SD. The SD network may comprise a U-Net architecture with an encoder and a (in particular skip-connected) decoder.

[0052] According to an example embodiment of the present invention, the diffusion model (also: diffusion probabilistic model, and/or score-based generative model) may comprise a machine learning (ML) model and in particular a generative model, which is configured for image generation, image denoising, inpainting, and/or super-resolution by means of a diffusion process, in particular by using a forward process (e.g., adding, in particular Gaussian, noise to an image), the reverse process (e.g., predicting a, in particular Gaussian, noise within the image and compensating and/or subtracting accordingly), and a sampling procedure.

[0053] According to an example embodiment of the present invention, SD may comprise a latent diffusion model. SD may relate to, or comprise, a neural network (NN) architecture (model) for text-to-image (in particular diffusion) generation. SD aims at learning a diffusion process that generates the probability distribution of a given (image) dataset.

[0054] According to an example embodiment of the present invention, the NN architecture of SD may comprise a convolutional NN (CNN)-based architecture, in particular a U-Net having an encoder and a decoder with skip connections (e.g., between layers of the encoder and the decoder) and/or a (e.g., cross-) attention mechanisms (e.g., within a layer of the encoder and/or the decoder). The NN architecture of SD may further comprise a diffusion process, by which an input image is transformed into a latent space image. By the U-Net, the latent space image may be transformed back into an image in non-latent (and/or image) space.

[0055] A denoising UNet may make use of latent diffusion, and/or may be trained in a latent space instead of an original image space. E.g., SD may be specifically trained in the latent space. Alternatively or in addition, generically diffusion models may be trained in the original (e.g., image) space.

[0056] According to an example embodiment of the present invention, training may comprise two stages with the first stage training an autoencoder to encode and/or transform an image in latent space and subsequently decode and/or reconstruct the image. In the second stage, the diffusion model may be trained in latent space with forward diffusion gradually adding (e.g., Gaussian) noise (in particular according to a Markov chain), and backward de-noising (and/or the UNet) predicting a (e.g., Gaussian) noise.

[0057] SD can advantageously produce high-quality (e.g., high-resolution and/or photorealistic) synthetic images based on (in particular visual domain and/or environmental class) textual prompts at low computational cost.

[0058] The computer-implemented technique of the present invention generalizes conventional generative image-to-feature models, in particular SD and/or VPD, by preserving generalizability and/or facilitating adapting to different visual domains, and optionally to different environmental classes, by a corresponding textual prompt.

[0059] Compared to generic “diffusion models”, SD may train a diffusion model in the latent space (in particular of a pretrained and/or, e.g., subsequently, frozen autoencoder), e.g., instead of in the image or pixel space. By training in the latent space, the computational cost can be reduced, as the latent space is much smaller than the conventional pixel space, e.g., eight (8) times smaller.

[0060] There exist alternatives to SD for the generative image-to-feature model (and/or the conditional feature extracting model), e.g., as described by Simian Luo et. al in “Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference”, arXiv:2310.04378v1 [cs.CV], which is incorporated herein by reference (and briefly denoted as LCM). In general, the technique for extracting features of an environment from image data is not limited to SD, but can, e.g., be performed using LCM. A main advantage of SD may be viewed in its open-sourced availability.

[0061] The conditioning model may comprise a ControlNet. The ControlNet may comprise an encoder and convolution layers with a cross-attention mechanism to the generative image-to-feature model, in particular to the encoder of the SD network.

[0062] According to an example embodiment of the present invention, the generative image-to-feature model, and in particular the SD network, can be conditioned by the visual domain by means of the conditioning model as an extra trainable branch, in particular by using the ControlNet. For the basics of conditioning a SD network using a ControlNet, it is referred to L. Zhang et al. in “Adding Conditional Control to Text-to-Image Diffusion Models” (2023), available at the Computer Vision Foundation and IEEE Xplore, which is incorporated herein by reference.

[0063] The SD may comprise a frozen (e.g., after an initial training) UNet, in particular to preserve trained (and/or prior) knowledge during fine-tuning. The ControlNet may comprise a trainable

copy of the UNet encoder of the SD for the conditioning by the visual domain textual prompt. The ControlNet may further comprise one or more zero-convolutional layers, through which the visual domain text prompt is fed into the decoder of the SD. A zero-convolutional layer may comprise a (e.g., 1×1) convolution layer with both weight and bias initialized as zero.

[0064] The connection between or combination of the SD network and the ControlNet can provide a particularly computationally efficient conditioning, which saves time and (e.g., graphics processing unit, GPU) memory.

[0065] The combination of the SD network and the ControlNet with the visual domain textual prompt, and optionally the environmental class textual prompt, can provide a particularly rich generalizable framework for feature extraction and/or visual perception task performance.

[0066] Conventional models for perception tasks are trained on a specific dataset, which may comprise limited data and/or scenarios, due to the costly annotation process (e.g., for a segmentation task). By contrast, the technique for extracting features of an environment from image data, e.g., according to the first method aspect of the present invention, leverages the rich prior knowledge of a pretrained large-scale text-to-image model, which has seen diverse scenarios and is equipped with a joint vision-language latent space. The technique for extracting features of an environment from image data in particular utilizes the language and/or textual prompt control potential and the powerful pretrained knowledge for improving the downstream task (e.g., AD, movement planning for a robot, and/or access control) model's generalization and adaptation capability in different operating domains during the deployment. Alternatively or in addition, only a lightweight decoder needs to be trained for the specific task, e.g., semantic segmentation, and/or object detection.

[0067] As to a second method aspect of the present invention, a computer-implemented method for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt is provided.

According to an example embodiment of the present invention, the method comprises a step of receiving a training dataset. The training dataset comprises a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment and (in particular local) features of the environment. The method further comprises a step of training a conditional feature extracting model based on the received training dataset. Training the conditional feature extracting model comprises receiving the image data of the environment as input, the visual domain textual prompt as condition, and the (in particular local) features, as ground truth.

[0068] The training dataset may comprise synthetic data (e.g., synthetic image data generated based on the, in particular local, features, and the visual domain textual prompt). Alternatively or in addition, the training dataset may comprise real data, in particular real image data (e.g., with the, in particular local, features and/or the visual domain prompt as annotation).

[0069] By training the conditional feature extracting model on a plurality of training datasets, in particular comprising synthetic training datasets, feature extraction, and/or performing a visual perception task, can be rendered adaptive (and/or generalizable) to (in particular previously untrained) visual domains and/or environmental classes, and/or to visual domains and/or environmental classes, for which real training datasets are scarce (e.g., due to real image data being scarce).

[0070] SD as an example may be pretrained on large-scale datasets, e.g., LAION-5B. The conditional feature extracting model can be trained on a small labeled real dataset for a specific task, e.g., semantic segmentation. For the AD use case, e.g., the Cityscapes Dataset (comprising 4,000 datasets with high quality annotation and 20,000 datasets with coarse annotation, www.cityscapes-dataset.com) can be used. Alternatively or in addition, synthetic data can also be used, in particular if the annotation (e.g., a segmentation label map) is generated along with the image as well.

[0071] The method according to the first method aspect of the present invention may be used for

autonomous driving (AD), planning a movement of a robot, operating a domestic appliance, and/or controlling an access control system.

[0072] By the (in particular local) feature extraction and/or by performing the visual perception task based on the image data and the visual domain textual prompt (and optionally the environmental class textual prompt), sufficiently accurate data on the environment may be provided to facilitate AD and/or path planning (and/or trajectory planning) for a (in particular AD) vehicle and/or a (in particular autonomous) robot and/or domestic appliance (e.g., an autonomously moving vacuum cleaner and/or artificial butler).

[0073] As to a first device aspect of the present invention, a computing device for extracting features of an environment from image data is provided. According to an example embodiment of the present invention, the computing device comprises a visual domain indication reception interface configured for receiving data indicative of a visual domain of an environment. The device further comprises a visual domain textual prompt generating module configured for generating a visual domain textual prompt based on the received data indicative of the visual domain of the environment. The computing device further comprises an environmental image data reception interface configured for receiving image data representative of the environment. The computing device still further comprises a conditional feature extracting model configured for extracting, in particular local, features of the environment from the received image data. The extracting of the (in particular local) features is conditioned by the generated visual domain textual prompt. Optionally, the computing device according to the first device aspect may comprise a visual perception task performing module configured for performing a visual perception task on the environment based on the extracted (in particular local) features. Further optionally, the computing device according to the first device aspect may comprise an environmental class receiving interface configured for receiving data indicative of a class in relation to the environment. The computing device may further comprise an environmental class textual prompt generating module configured for generating a environmental class textual prompt based on the received data indicative of the class in relation to the environment. Extracting (in particular local) features of the environment from the received image data may be further conditioned by the generated environmental class textual prompt.

[0074] The computing device according to the first device aspect of the present invention may be configured to perform any one of the steps, and/or comprise any one of the features, disclosed in the context of the first method aspect.

[0075] As to a second device aspect of the present invention, a computing device for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt is provided. The computing device comprises a training data reception interface configured for receiving a training dataset. The training dataset comprises a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment and (in particular local) features of the environment. The computing device further comprises a training module configured for training a conditional feature extracting model based on the received training dataset. Training the conditional feature extracting model comprises receiving the image data of the environment as input, the visual domain textual prompt as condition, and the (in particular local) features as ground truth.

[0076] The computing device according to the second device aspect of the present invention may be configured to perform any one of the steps, and/or comprise any one of the features, disclosed in the context of the second method aspect.

[0077] The computing devices according to the first and the second method aspect of the present invention may be identical.

[0078] As to a system aspect of the present invention, a system for extracting features of an environment from image data is provided. According to an example embodiment of the present invention, the system comprises the computing device according to the first device aspect and at

least one sensor, and/or at least one interface, for receiving the data indicative of the visual domain and the image data representative of the environment. The data indicative of the visual domain and/or the image data representative of the environment may be received wirelessly (e.g., from a sensor and/or webserver outside of an AD vehicle). Alternatively or in addition, the data indicative of the visual domain and/or the image data representative of the environment may be received via a wired connection (e.g., for a fixedly installed access control system).

[0079] As to a further aspect of the present invention, a computer program product is provided. The computer program product comprises program elements which induce a computing device to carry out the steps of the method for extracting features of an environment from image data, and/or for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt, according to the first and/or second method aspect, when the program elements are loaded into a memory of the computing device.

[0080] As to a still further aspect of the present invention, a computer-readable medium is provided, on which program elements are stored that can be read and executed by a computing device, in order to perform steps of the method for extracting features of an environment from image data, and/or for training a conditional feature extracting model for extracting, in particular local, features of an environment from image data conditioned by a visual domain textual prompt, according to first and/or second method aspect, when the program elements are executed by the computing device.

Description

BRIEF DESCRIPTION OF THE DRAWINGS

[0081] FIG. 1 is a flowchart of a method for extracting features of an environment from image data, according to an example embodiment of the present invention.

[0082] FIG. 2 is a flowchart of a method for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt, according to an example embodiment of the present invention.

[0083] FIG. 3 is a schematic illustration of an architecture of a computing device for extracting features of an environment from image data, according to an example embodiment of the present invention.

[0084] FIG. 4 is a schematic illustration of an architecture of a computing device for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt, according to an example embodiment of the present invention.

[0085] FIG. 5 schematically illustrates the conventional SD architecture.

[0086] FIG. 6 schematically illustrates the conventional VPD architecture.

[0087] FIG. 7 schematically illustrates the conventional ControlNet architecture.

[0088] FIG. 8 schematically illustrates the technique for extracting features of an environment from image data.

[0089] FIGS. 9A and 9B exemplarily illustrate data indicative of a visual domain of an environment along with a probability analysis of a predetermined set for generating (in particular automatically) the visual domain textual prompt.

[0090] FIGS. 10A, 10B and 10C schematically illustrate use cases of the technique for extracting features of an environment from image data, in particular for automated driving, planning of a movement of a robot, and access control, respectively.

DETAILED DESCRIPTION OF EXAMPLE EMBODIMENTS

[0091] FIG. 1 schematically illustrates a computer-implemented method **100** for extracting features

of an environment from image data.

[0092] The method **100** comprises a step **S102** of receiving data indicative of a visual domain of an environment. The method **100** further comprises a step **S104** of generating a visual domain textual prompt based on the received **S102** data indicative of the visual domain of the environment. In a step **S106**, image data representative of the environment are received. The method further comprises a step **S108** of extracting (in particular local) features of the environment from the received **S106** image data. The extracting **S108** of the (in particular local) features is performed by a conditional feature extracting model. The extracting **S108** of the (in particular local) features is conditioned by the generated **S104** visual domain textual prompt.

[0093] Optionally, in a step **S110** a visual perception task on the environment is performed based on the extracted **S108** (in particular local) features.

[0094] Further optionally, the method **100** comprises a step **S101** of receiving data indicative of an environmental class in relation to the environment and a step **S103** of generating an environmental class textual prompt based on the received **S101** data indicative of the class in relation to the environment. Extracting **S108** (in particular local) features of the environment from the received **S106** image data may be further conditioned by the generated **S103** environmental class textual prompt.

[0095] FIG. 2 schematically illustrates a computer-implemented method **200** for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt.

[0096] The method **200** comprises a step **S202** of receiving a training dataset. The training dataset comprises a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment and (in particular local) features of the environment. The method **200** further comprises a step **S204** of training a conditional feature extracting model based on the received **S202** training dataset. Training **S204** the conditional feature extracting model comprises receiving the image data of the environment as input, the visual domain textual prompt as condition, and the (in particular local) features as ground truth.

[0097] FIG. 3 schematically illustrates an architecture of a computing device **300** for extracting features of an environment from image data.

[0098] The computing device **300** comprises a visual domain indication reception interface **302** configured for receiving data indicative of a visual domain of an environment. The computing device **300** further comprises a visual domain textual prompt generating module **304** configured for generating a visual domain textual prompt based on the received data indicative of the visual domain of the environment. The computing device **300** further comprises an environmental image data reception interface **306** configured for receiving image data representative of the environment. The computing device **300** still further comprises a conditional feature extracting model **308** configured for extracting (in particular local) features of the environment from the received image data. The extracting of the (in particular local) features is conditioned by the generated visual domain textual prompt.

[0099] Optionally, the computing device **300** may comprise a visual perception task performing module **310** configured for performing a visual perception task on the environment based on the extracted (in particular local) features.

[0100] Further optionally, the computing device **300** may comprise an environmental class receiving interface **301** configured for receiving data indicative of a class in relation to the environment. The computing device may further comprise an environmental class textual prompt generating module **303** configured for generating an environmental class textual prompt based on the received data indicative of the class in relation to the environment. Extracting (in particular local) features of the environment from the received image data may be further conditioned by the generated environmental class textual prompt.

[0101] Any one of the visual domain indication reception interface **302**, the optional environmental

class receiving interface **301**, and the environmental image data reception interface **306** may be embodied by an input-output interface **312**. Alternatively or in addition, any one of the visual domain textual prompt generating module **304**, the optional environmental class textual prompt generating module **303**, the conditional feature extracting model **308**, and the optional visual perception task performing module **310** may be embodied by a processing unit **314**. Further alternatively or in addition, the computing device **300** may comprise at least one memory **316**. [0102] The computing device **300** may be configured to perform any one of the steps, and/or comprise any one of the features, disclosed in the context of the method **100**.

[0103] FIG. **4** schematically illustrates an architecture of a computing device **400** for training a conditional feature extracting model for extracting (in particular local) features of an environment from image data conditioned by a visual domain textual prompt.

[0104] The computing device **400** comprises a training data reception interface **402** configured for receiving a training dataset. The training dataset comprises a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment and, in particular local, features of the environment. The computing device **400** further comprises a training module **404** configured for training a conditional feature extracting model based on the received training dataset. Training the conditional feature extracting model comprises receiving the image data of the environment as input, the visual domain textual prompt as condition, and the (in particular local) features as ground truth.

[0105] The training data reception interface **402** may be embodied by an input-output interface **406**. Alternatively or in addition, the training module **404** may be embodied by a processing unit. Further alternatively or in addition, the computing device **400** may comprise at least one memory **410**.

[0106] The computing device **400** may be configured to perform any one of the steps, and/or comprise any one of the features, disclosed in the context of the method **200**.

[0107] Any one of the processing units **314**; **408** may comprise a central processing unit (CPU), and/or a graphics processing unit (GPU).

[0108] The computing device **300** and the computing device **400** may be identical.

[0109] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data provide a novel way of utilizing a large-scale text-to-image model (in particular the conditional feature extracting model) for domain adaptive visual perception tasks, e.g., semantic segmentation and/or object detection.

[0110] Prior work on VPD making use of finetuned SD and a trained task-specific decoder has demonstrated promising results when testing on the same trained domain. The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data push one step further and aim to employ the model on various unseen domains, taking advantage of the generalizable representation of the conditional feature extracting model (and/or large-scale pretrained model) and the joint vision-language embedding space. The (in particular visual domain and/or environmental class) textual prompts are leveraged, e.g., as the domain indicator to adapt the model to different domains. The conditional feature extracting model (as a text-to-image diffusion model) can (e.g., naturally) be conditioned on the text. Additionally, a task-specific decoder can be conditioned on the domain embeddings as well to further enhance the conditioning, as schematically illustrated in FIG. **8**.

[0111] E.g., SD not only captures rich semantic information, but may also has seen many visual domains, e.g., different weather and/or lightning conditions, making its representation more generalizable. Differently from VPD, which potentially overfits on the training set, the generalizable representation (e.g., of SD) can be leveraged according to the techniques for extracting features of an environment from image data and for training a conditional feature

extracting model for extracting (in particular local) features of the environment from the image data and employed for domain adaptation and/or generalization tasks. E.g., SD may be finetuned on the training set while preserving its ability to be adapted to different test sets. For instance, the conditional feature extracting model can be finetuned on sunny images, and the conditional feature extracting model will have reasonable performance not only on the trained domain (e.g., sunny condition), but also on unseen data, e.g., rainy and/or foggy image data.

[0112] It is note that the conventional VPD finetunes the whole denoising model directly, which might destroy the prior knowledge, as illustrated in FIG. 6.

[0113] As an alternative, the fine-tuning manner of the ControlNet may be adopted, which adds a trainable branch while freezing the original SD to maximally preserve the prior knowledge, as schematically illustrated in FIG. 7. Additionally, the techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data exploit the usage of (e.g., visual domain and/or environmental class) textual prompts, in particular as a domain indicator, to conveniently adjust the a conditional feature extracting model to different domains, leveraging the joint visual and text embedding space of SD.

[0114] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data are more robust and generalizable to various (in particular unseen) visual domains, e.g., different weather and/or lightning condition, and/or new countries. Conventionally, the network tends to overfit to the training set (e.g., sunny images), and cannot generalize to unseen domains (e.g., rainy and/or foggy scenes). Since the environment in the real-world is diverse and changing, and the conditional feature extracting model can only be trained on limited data, it is important to design a robust conditional feature extracting model that can cope with unseen situations. This is critical for real-life applications such as AD (in particular autonomous driving), and/or robots. For instance, a person walking on the road at night could be ignored by an autonomous car and lead to an accident, if the conditional feature extracting model trained on sunny images cannot generalize to night scenes.

[0115] By contrast, the technique for extracting features of an environment from image data can adapt itself in different domains flexibly with the aid of the (e.g., visual domain and/or environmental class) textual prompt. Alternatively or in addition, the conditional feature extracting model is built on top of a large-scale pretrained model (e.g., SD), which can transfer rich prior knowledge and make the training converge faster.

[0116] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data can be used for analyzing data obtained from a sensor. The sensor may determine measurements of the environment in the form of sensor signals, which may be given by (in particular digital) image data, comprising e.g. video, radar, LiDAR, ultrasonic, motion, and/or thermal images.

[0117] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data can be used for classifying the sensor data, detecting the presence of objects in the sensor data, and/or performing a semantic segmentation on the sensor data, e.g., regarding traffic signs, road surfaces, pedestrians vehicles, and/or object classes that could show in the semantic segmentation task, such as trees and/or the sky.

[0118] Alternatively or in addition, the techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data can be used for determining a continuous value or multiple continuous values (e.g., perform a regression analysis), e.g., regarding a distance, a velocity, an acceleration, and/or tracking an item, e.g., an object, in the data. The determining of the

one or more continuous values may be carried out based on low-level features (e.g. edges and/or pixel attributes for images).

[0119] Alternatively or in addition, the techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data can be used to detect anomalies in a (in particular downstream) technical system (e.g., an AD vehicle, a robot, a domestic appliance, and/or an access control system). It can do so by analyzing data (e.g. scalar time series, also denoted as measure and control), especially from a sensor and then operating the technical system accordingly.

Alternatively or in addition, anomalies may be detected and then the technical system may be operated in a fail-safe mode.

[0120] Method/Data for Training: The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data may be considered as the upstream part in a machine learning (ML) tool chain. After being trained, the conditional feature extracting model of ML system can be put to downstream use, e.g., for AD, planning of movements of a robot, operating of a domestic appliance, and/or or controlling an access control system.

[0121] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data according to an embodiment employs SD not only on the trained domain, but also preserves its generalization ability so that it can be used in unseen domains as well. Thanks to the joint vision-language embedding space of SD, the conditional feature extracting model can adapt (e.g., to a different domain) using the (in particular visual domain and/or environmental class) textual prompt.

[0122] As already mentioned above, instead of operating in the image space, SD operates in the latent space of an autoencoder, as schematically illustrated in FIG. 5. Firstly, the encoder ϵ maps the given image x into a spatial latent code $z=\epsilon(x)$. Then z is mapped back to the image space by the decoder D .

[0123] The autoencoder is trained to reconstruct the given image, i.e., $D(\epsilon(x))\approx x$.

[0124] In the second stage, a diffusion model is trained in the latent space Z . The diffusion model

[0125] comprises a forward diffusion process and a backward denoising process. The forward

[0126] pass comprises a Markov chain to gradually add Gaussian noise to the clean data. Formally, it can be written as, e.g., $q(z_{\text{sub}.t}|z_{\text{sub}.t-1})=N(z_{\text{sub}.t};\sqrt{\text{square root over } (1-\beta_{\text{sub}.t})}z_{\text{sub}.t-1}, \beta_{\text{sub}.t})$, where $\{\beta_{\text{sub}.t}\}_{\text{sub}.t=0}^{\text{sup}.T}$ are fixed variance schedule.

[0127] The noisy latent can be computed in a closed form, e.g.,

[00001] $z_t = \sqrt{t}z_0 + \sqrt{1 - t} \cdot \tilde{N}(0, I)$, (1) [0128] where $z_{\text{sub}.0}=\epsilon(x)$, and

$\alpha_{\text{sub}.t}=\prod_{\text{sub}.s=1}^{\text{sup}.t}(1-\beta_{\text{sub}.s})$. The reverse denoising process can be parametrized by another

Gaussian distribution $p_{\text{sub}.\theta}(z_{\text{sub}.t-1}|z_{\text{sub}.t}):=N(z_{\text{sub}.t-1}; \mu_{\text{sub}.\theta}(z_{\text{sub}.t}, t), \sigma_{\text{sub}.\theta}(z_{\text{sub}.t}, t))$.

Essentially, $\mu_{\text{sub}.\theta}(z_{\text{sub}.t}, t)$ can be expressed as a linear combination of $z_{\text{sub}.t}$ and the predicted noise $\epsilon_{\text{sub}.\theta}(z_{\text{sub}.t}, t)$, which is modelled by a Unet. The parameters of the Unet may be learned by minimizing the L2 norm of the noise prediction at a sampled timestep t :

[00002] $\mathcal{L}_{\text{noise}} = E_{z \sim (x), \tilde{N}(0, I), t} [\text{.Math.} - (z_t, t) \text{.Math.}]^2$. (2)

[0129] A more mathematical derivation is provided by Jonathan Ho et al. in “Denoising Diffusion Probabilistic Models”, arXiv:2006.11239v2 [cs.LG], which incorporated herein by reference. The conditional information, e.g., (in particular visual domain and/or environmental class) text prompt, is inserted via the cross-attention mechanism indicated a “CLIPtext” in the lower part of FIG. 5.

[0130] At inference time, the latent representation Z_r may be randomly sampled from the Gaussian distribution, and then the trained denoising Unet may be employed sequentially, to obtain the denoised latent $z_{\text{sub}.t-1}$ given $z_{\text{sub}.t}$ from $t=T$ to $t=1$. The final synthesized image may be obtained by feeding the clean latent $z_{\text{sub}.0}$ through the decoder D .

[0131] VPD, as schematically indicated in FIG. 6 by the flame symbol, finetunes the pre-trained SD Unet for downstream visual perception tasks, e.g., semantic segmentation and/or depth estimation. As schematically illustrated in FIG. 6, VPD extracts features and cross-attention maps from the Unet and feeds them to the task specific decoder. For a semantic segmentation task, the class embedding can be extracted from a frozen (as indicated by the snowflake symbol) CLIP text encoder and adapted by a trainable text adapter (as indicated by the flame symbol, e.g., a few MLP layers), which can then be used as the (e.g., domain and/or class) textual prompt (also: text condition) to the SD Unet. Nevertheless, VPD only focuses on the in-domain ability, e.g., the test set is assumed to share the same distribution as the training set, and the large model tends to overfit to the specific trained domain.

[0132] By contrast, the techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data take the conditional feature extracting model's generalization ability into consideration and design a more robust conditional feature extracting model also for unseen domains and/or unseen environments.

[0133] The ControlNet enhances the SD with additional input conditions, e.g., a label map. As schematically illustrated in FIG. 7, the ControlNet clones a trainable copy of the Unet encoder, and the original SD is frozen during the fine-tuning (as indicated by the flame and snowflake symbols, respectively). The input condition is fed as the input of the ControlNet, and after the zero convolution layers and trainable encoder, the features are inserted back to the decoder of SD, where the zero convolution layers are essentially 1×1 convolution layers with both weight and bias initialized as zeros.

[0134] The training objective of the ControlNet plus SD Unet architecture is adopted from eq. (2) by inserting the condition y :

$$[00003] \mathcal{L}_{\text{noise}} = E_{z \sim (x), \sim N(0, I), t} [\cdot \text{Math.} - (z_t, t, y) \cdot \text{Math.}]^2. \quad (3)$$

[0135] The techniques for extracting features of an environment from image data and for training a conditional feature extracting model for extracting (in particular local) features of the environment from the image data exploit the conditional feature extracting model, in particular as a large-scale text-to-image generative model (e.g., comprising SD) for visual perception tasks (e.g., semantic segmentation and/or object detection), while preserving the conditional feature extracting model's generalization ability in various (in particular unseen) domains and/or environments. The key idea is to take advantage of the joint vision-language latent space of large-scale models and design the conditional feature extracting model to adapt itself flexibly based on the (in particular visual domain and/or environmental class) textual prompt (also: text condition) indicating the visual domain and/or the environmental class.

[0136] As indicated at reference signs **802** and **804**, given data indicative a class in relation to an environment and/or indicative of a visual domain of the environment (e.g, comprising an image), the operating visual domain and/or environmental class that it belongs to is determined. This can be done either manually or automatically. For the manual case, the operating “visual domain” and/or environmental class can be specified directly by human (e.g., via a UI, in particular a GUI).

[0137] Alternatively or in addition, some datasets such as the Adverse Conditions Dataset (ACDC, available at <https://acdc.vision.ee.ethz.ch/>), already comprise the meta-information of weather conditions (e.g., rain, fog, snow, and/or night). In downstream (and/or real) applications, such as AD (in particular autonomous driving), a human can provide such information to the model dynamically.

[0138] For the automatic identification of the operating visual domain and/or environmental class, a pretrained Visual Question Answering (VQA) model, e.g., mPlug, may be leveraged (e.g., as described by Chenliang Li et al., “mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections”, arXiv:2205.12005v2 [cs. CL] and incorporated herein by

reference). Examples are shown in FIGS. 9A and 9B, where a model asks “What weather condition is it?” based on an image **804** as the data indicative of the visual domain. The model returns top answers 812 with likelihood. E.g., in FIG. 9A both “sunny” and “cloudy” are the two most likely answers with nearly identical probability, whereas in FIG. 9B, “cloudy” is the answer with the by far largest probability.

[0139] Similarly, questions may be asked, e.g., with respect to lightning conditions, and/or countries. The top-1 answer can be taken as the (e.g., intermediary) visual domain prompt input and/or data indicative of the visual domain (e.g., for the example in FIG. 9B). Alternatively or in addition, as the “visual domain” concept may be ambiguous (e.g., with nearly equal probabilities, as exemplified in FIG. 9A), the top-K answers may be taken with probability (e.g., with K=2). Having the (e.g., intermediary) visual domain prompt input and/or data indicative of the visual domain, a frozen CLIP encoder **806** and a (in particular light-weight) trainable text adapter **808** may be employed to obtain the visual domain textual prompt (also denoted as domain embeddings $W_{\text{sub.i}}$) at reference sign **812**, e.g.:

$$[00004] \quad W_i = \text{adp}(\text{CLIP}_{\text{text}}(y_i)), i = 1, \dots, K \quad (4)$$

where $y_{\text{sub.i}}$ represents the visual domain prompt, e.g., “sunny scenes”, and $\text{CLIP}_{\text{sub.text}}$ and $\psi_{\text{sub.adp}}$ are CLIP text encoder and the text adapter, respectively.

[0140] As schematically illustrated in FIG. 8, the conditional feature extracting model may be further conditioned on an environmental class textual prompt (also denoted as class embedding c) at reference sign **810**, which may be obtained from data indicative of the class in relation to the environment **802** using a further frozen CLIP encoder **806** and a further a (in particular light-weight) trainable text adapter **808**.

[0141] The visual domain textual prompt (and/or domain embeddings) **812** may be used to condition the conditional feature extracting model comprising the exemplary frozen SD Unet **822** and the ControlNet **814** having a trainable copy **816** of the Unet encoder (e.g., of the SD Unet **822**) followed by zero convolutions **818**. As an alternative option of direct fine-tuning the whole Unet as in conventional VPD, in the example of FIG. 8, the fine-tuning manner of ControlNet **814**, which keeps the original Unet **822** frozen and adds the trainable copy **816** of the Unet encoder, which is initialized using the pretrained weights of the SD Unet encoder. In this way, the prior knowledge of SD is fully preserved.

[0142] Both the frozen SD Unet **822** and the trainable copy **816** of the Unet encoder are configured to receive the latent representation **820**.

[0143] Naturally, the optional environmental class textual prompt **810** (also: text condition) may be used in the same way as in conventional SD in FIG. 8, which can be utilized via cross-attention. The visual domain textual prompt **812** (also: domain embeddings) in the exemplary embodiment of FIG. 8 is fed into the trainable UNet encoder **816** of the ControlNet **814** using a cross-attention mechanism.

[0144] The features and cross-attention maps can be extracted analogously to VPD, as schematically illustrated at reference sign **826**. Alternatively or in addition, instead of extracting from the single clean latent $z_{\text{sub.0}}$, the features may be extracted from multiple timesteps. Based on eq. (1), noisy latents may be obtained at different timesteps directly. The feature extraction step may, e.g., be formulated as:

$$[00005] \quad f_{i,t} = \text{SD}(W_i, z_t, c), f_i = \text{Agg}(f_{i,t}), f = \text{Math.}_i f_i * p_i, \quad (5)$$

where SD represents Stable Diffusion, and c is the class embeddings (and/or environmental class textual prompt) **810** for the task, e.g., semantic segmentation. Agg indicates the aggregation operation, e.g., a simple mean average over features $f_{\text{sub.i,t}}$ at different timesteps. $p_{\text{sub.i}}$ is the probability of the current domain embedding $W_{\text{sub.i}}$ (and/or visual domain textual prompt) **812**.

[0145] Further, the visual domain textual prompt (and/or domain embeddings) **812** may be used in the task specific decoder **828**. The domain embedding (visual domain textual prompt) **812** may be

firstly aggregated as indicated at reference sign **824**,

$$[00006] W_{agg} = \text{Agg}(W_i, p_i), \quad (6)$$

where Agg is the aggregation operation, e.g., a simple probabilistic average and/or learnable [0146] mapping. As mentioned in previous works, e.g., by Seonguk et al. in “Learning to Optimize Domain Specific Normalization for Domain Generalization”, DOI: 10.1007/978-3-030-58542-6_5, which is incorporated herein by reference, normalization statistics can have important impact in a visual domain (and/or environmental class) generalization ability. Therefore, according to an embodiment, the domain embedding (and/or visual domain textual prompt) **812** is used to map these embeddings to the normalization parameters, e.g.:

$$[00007] f_{adp} = \frac{f_{fdec} - \mu_{fdec}}{\sigma_{fdec}} (W_{agg}) + \gamma_{fdec} (W_{agg}), \quad (7)$$

where μ_{fdec} and σ_{fdec} are the mean and standard deviation of features within the decoder f_{subdec} (e.g., the task-specific decoder **828**). $\mu_{W_{agg}}$ and $\gamma_{W_{agg}}$ are the learned affine transformation parameters, conditioned on the aggregated domain embedding W_{subagg} . Finally, the task specific decoder **828** will output the predictions on the specific task (e.g., of the semantic segmentation and/or object classification).

[0147] FIG. **10A** schematically illustrates an example of a first use case, in which the AD vehicle's **1002-1** controller **1004** receives the result of the feature extraction (e.g., according to the step **S108**) and/or of the visual perception task (e.g., according to the step **S110**) for performing the AD functionality. Optionally, the received result may also displayed to a driver and/or passenger, e.g., on a display **1006** of a center console.

[0148] FIG. **10B** schematically illustrates an example of a second use case, in which a robot's **1002-2** controller **1004** receives the result of the feature extraction (e.g., according to the step **S108**) and/or of the visual perception task (e.g., according to the step **S110**) for planning the motion of the robot **1002**.

[0149] FIG. **10C** schematically illustrates an example of a third use case, in which an access control system **1002-3** has a first sensor **1010** for acquiring the image data representative of the environment, optionally a second sensor **1012** for acquiring audio data representative of the environment, a controller **1004** and optionally a display **1006** (in particular for a person requesting access). The controller **1004** receives the result of the feature extraction (e.g., according to the step **S108**) and/or of the visual perception task (e.g., according to the step **S110**) for determining if access is granted or not.

Cited Related Art

[0150] [1] R. Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, arXiv:2112.10752v2 [cs.CV] [0151] [2] Wenliang Zhao et al., “Unleashing Text-to-Image Diffusion Models for Visual Perception, arXiv:2303.02153 v1 [cs.CV] [0152] [3] P. Esser et al., “Taming Transformers for High-Resolution Image Synthesis”, arXiv:2012.09841v3 [cs.CV] [0153] [4] Yong Zheng Ong et al., “VG-GAN: Conditional GAN Framework for Graphical Design Generation”, 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 781-785, doi: 10.1109/ICIP46576.2022.9897935 [0154] [5] Dan Zhang et al., “Anomaly-Aware Semantic Segmentation via Style-Aligned OoD Augmentation”, arXiv:2308.09965v1 [cs.CV] [0155] [6] Simian Luo et al., “Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference”, arXiv:2310.04378v1 [cs.CV] [0156] [7] www.cityscapes-dataset.com/ [0157] [8] Jonathan Ho et al., “Denoising Diffusion Probabilistic Models”, arXiv:2006.11239v2 [cs. LG] [0158] [9] acdc.vision.ee.ethz.ch/ [0159] [10] Chenliang Li et al., “mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections”, arXiv:2205.12005v2 [cs. CL] [0160] [11] Seonguk Seo et al., “Learning to Optimized Domain Specific Normalization for Domain Generalization” (2020) 10.1007/978-3-030-58542-6_5.

Claims

1. A computer-implemented method for extracting features of an environment from image data, comprising the following steps: receiving data indicative of a visual domain of an environment; generating a visual domain textual prompt based on the received data indicative of the visual domain of the environment; receiving image data representative of the environment; and extracting local features of the environment from the received image data, wherein the extracting of the local features is performed by a conditional feature extracting model, and wherein the extracting of the local features is conditioned by the generated visual domain textual prompt.
2. The method according to claim 1, further comprising: performing a visual perception task on the received image data based on the extracted local features.
3. The method according to claim 2, wherein the visual perception task includes at least one of: a depth estimation; an object detection; a classification; a semantic segmentation.
4. The method according to claim 2, wherein the performing of the visual perception task on the image data is conditioned on the generated visual domain textual prompt.
5. The method according to claim 1, wherein the received data indicative of the visual domain include at least one of: environmental sensor data; position data determined using a positioning system, including a satellite navigation system; manually input data; electronically available information data.
6. The method according to claim 1, further comprising: receiving data indicative of a class in relation to the environment; and generating an environmental class textual prompt based on the received data indicative of the class in relation to the environment; wherein the extracting of the local features of the environment from the received image data is further conditioned by the generated environmental class textual prompt.
7. The method according to claim 6, wherein the extracting of the local features is further conditioned by the environmental class textual prompt.
8. The method according to claim 6, wherein the generating of the visual domain textual prompt, and/or the generating of the environmental class textual prompt is performed by an adaptive contrastive language-image pretraining (CLIP) encoder.
9. The method according to claim 1, wherein the conditional feature extracting model includes a generative image-to-feature model, which is configured to extract the local features of the environment from the received image data of the environment, and a conditioning model, which is configured to encode the generated visual domain textual prompt for conditioning, and/or controlling, the generative image-to-feature model.
10. The method according to claim 9, wherein the generative image-to-feature model includes a diffusion model including Stable Diffusion (SD), wherein the SD includes a U-Net architecture with an encoder and a skip-connected decoder.
11. The method according to claim 9, wherein the conditioning model includes a ControlNet, wherein the ControlNet includes an encoder and convolution layers with a cross-attention mechanism to the generative image-to-feature model.
12. A computer-implemented method for training a conditional feature extracting model for extracting local features of an environment from image data conditioned by a visual domain textual prompt, comprising the following steps: receiving a training dataset, wherein the training dataset includes a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment and local features of the environment; and training a conditional feature extracting model based on the received training dataset, wherein the training of the conditional feature extracting model includes receiving the image data of the environment as input, the visual domain textual prompt as a condition, and the local features as ground truth.
13. The method according to claim 1, wherein the method is for at least one of the following:

autonomous driving; planning a movement of a robot; operating a domestic appliance; controlling an access control system.

14. A computing device configured to extract features of an environment from image data, the computing device comprising: a visual domain indication reception interface configured for receiving data indicative of a visual domain of an environment; a visual domain textual prompt generating module configured to generate a visual domain textual prompt based on the received data indicative of the visual domain of the environment; an environmental image data reception interface configured to receive image data representative of the environment; and a conditional feature extracting model configured to extract local features of the environment from the received image data, wherein the extracting of the local features is conditioned by the generated visual domain textual prompt.

15. A computing device configured to training a conditional feature extracting model to extract local features of an environment from image data conditioned by a visual domain textual prompt, the computing device comprising: a training data reception interface configured to receive a training dataset, wherein the training dataset includes a visual domain textual prompt indicative of a visual domain of an environment, image data of the environment, and local features of the environment; and a training module configured to train a conditional feature extracting model based on the received training dataset, wherein training the conditional feature extracting model includes receiving the image data of the environment as input, the visual domain textual prompt as condition, and the local features as ground truth.
