InAction: Interpretable Action Decision Making for Autonomous Driving

Taotao Jing[1], Haifeng Xia[1], Renran Tian[2], Haoran Ding[2], Xiao Luo[2], Joshua Domeyer[3], Rini Sherony[3], and Zhengming Ding[1]

[1] Tulane University, New Orleans LA 70118, USA
[2] Indiana University-Purdue University Indianapolis , Indianapolis IN 46202, USA
[3] Collaborative Safety Research Center (CSRC), Toyota Motor North America, Ann Arbor MI 48105, USA
{tjing, hxia, zding1}@tulane.edu, {rtian, luo25}@iupui.edu, {joshua.domeyer, rini.sherony}@toyota.com, hd10@iu.edu

Abstract. Autonomous driving has attracted interest for interpretable action decision models that mimic human cognition. Existing interpretable autonomous driving models explore static human explanations, which ignore the implicit visual semantics that are not explicitly annotated or even consistent across annotators. In this paper, we propose a novel Interpretable Action decision making (**InAction**) model to provide an enriched explanation from both explicit human annotation and implicit visual semantics. First, a proposed visual-semantic module captures the region-based action-inducing components from the visual inputs, which learns the implicit visual semantics to provide a human-understandable explanation in action decision making. Second, an explicit reasoning module is developed by incorporating global visual features and action-inducing visual semantics, which aims to jointly align the human-annotated explanation and action decision making. Experimental results on two autonomous driving benchmarks demonstrate the effectiveness of our **InAction** model for explaining both implicitly and explicitly by comparing it to existing interpretable autonomous driving models. The source code is available at https://github.com/scottjingtt/InAction.git.

Keywords: interpretable machine learning, action decision prediction

1 Introduction

Deep learning has recently accelerated the progress of autonomous through remarkable success in computer vision tasks. Existing driving action decision systems can primarily be recognized to be in two major groups, one is the *pipelined* framework [41] and the other is *end-to-end* system [17], [38], [18], [34], [35], [33]. Specifically, pipelined systems decompose the problem into a series of smaller tasks, such as pedestrian trajectory planning and object detection. The final driving action decision is made by relying on the performance of all the modules designed for the sub-tasks. However, pipelined systems are vulnerable to inaccuracies in each sub-task module, which may cause the entire system to perform

unreliably if the interactions between modules are ignored. On the contrary, end-to-end systems take advantage of the entire visual scene to directly predict driving action, avoiding the loss of information caused by the intermediate decisions adopted in pipelined systems.

Unfortunately, most end-to-end systems are complex deep neural network models, performing as a black box with opaque reasoning for human interpretation. In safety-critical domains, such as autonomous driving and medical diagnosis, building a transparent and interpretable learning model has recently attracted attention beyond the performance alone [28]. Various interpretation strategies have been explored to explain learning models, e.g., part-based methods [45], [47], saliency maps [1], [12], [46], activation maximization to visualize neurons [23], [24], deconvolution/upconvolution to explain layers [10], [42]. However, such post-hoc methods give a superficial understanding of the black box models, rather than being a comprehensive explainable system [28]. Alternatively, prototypical visual explanations are incorporated in deep network architecture for intrinsic interpretation and case-based reasoning [2], [29], [22], [21]. Most prior prototype-based work explicitly explores the presence of prototypical parts, which are utilized to recognize objects. However, such strategies ignore the notion of spatial relationships, which is crucial for tasks like driving decision making with complicated context and multiple objects.

For explainable autonomous driving decision making, Xu *et al.* proposed a new paradigm to predict driving action based on finite action-inducing objects, and generated a set of potential explanations in a multi-task fashion [39]. Unfortunately, there are four major limitations of this work from an interpretability perspective. First, although the multi-task framework is supervised by both driving action and a human-defined explanation, the proposed model does not interpret the reasoning process of the prediction for black-box model. Second, the proposed BDD-OIA dataset annotates the reasons of action into 21 explanations; however, it is impractical that the human-defined finite explanation set can cover all possible scenarios considering the complex scene context and objects input for autonomous driving action prediction tasks. For example, the explanation set in the BDD-OIA dataset recognizes "obstacles on the right lane" as a reason "cannot turn right", which is not accurate since different distances and locations of the obstacles could lead to different decisions for drivers. Moreover, the logical reasoning process from the explanation to driving action decision is ambiguous, especially under a multi-label setting that all possible actions are annotated. For instance, we notice that the proposed model predicts two explanations "traffic light is green" and "obstacle: car", but still predicts the action as "forward", without any reasoning about how the predicted explanation results in the action prediction. Last but not least, OIA estimates the driving decision only based on the last frame of observed sequence, ignoring the temporal information.

In this paper, we propose a novel Interpretable Action decision making (**InAction**) to provide reasoning of action prediction from both explicit human annotation and implicit visual semantics (Figure 1). Generally, we consider the explanation for action decision from two perspectives to compensate for the lim-

itations of each method: existing human-annotated interpretation and AI-based implicit visual hints. To sum up, our contributions are in three areas:

- First, we propose an inherently interpretable reasoning framework for autonomous driving action prediction from both implicit visual semantics and explicit human annotation perspectives.
- Second, the proposed *Implicit Visual-Semantic Interpretation* module interacts with the *Explicit Human-like Reasoning* module by revealing action-inducing concepts, and the learned implicit and explicit explanations compensate for the limitations of each other in predicting the action decision.
- Finally, experimental results on two explainable autonomous driving benchmarks demonstrate the effectiveness of the proposed model by comparing with existing models showing enriched interpretation and reasoning.

2 Related Work

2.1 Autonomous Driving Action Prediction

Existing autonomous driving action prediction solutions can be roughly grouped into two branches, i.e., *end-to-end* and *pipelined*. Generally, pipelined frameworks separate the problem into a series of smaller tasks, such as object detection [6], [3],[15], pedestrian trajectory planning [30], [26], [5], [31], scene segmentation [44], [11], [32], and object tracking [19], [7], [20]. Exploring the performance of each sub-module and assessing its potential contribution to the final prediction can help users understand the reasoning of the final decision. However, because the final decision prediction relies on the performance of all sub-task modules, pipelined systems are vulnerable to failures of individual sub-task modules, leading to unreliable systems.

End-to-end autonomous driving systems have achieved promising progress thanks to the success of computer vision deep leaning algorithms [17], [38], [18], [34], [35], [33],[9], [18], [37], [36]. However, most end-to-end systems are complex deep neural networks, requiring large-scale datasets to train. Unfortunately, the black box nature of deep neural networks makes the decisions not always trustworthy. Xu *et al.*. design an explainable object-inducing driving action prediction system (OIA) together with a new benchmark consisting of ego-view driving videos annotated with action and static explanations [39]. Through a model that jointly predicts action and explanation, the learned model can explain the decision predicted within the pre-defined reasoning set.

2.2 Interpretable Machine Learning

Building a transparent and interpretable model is crucial for safety-critical problems, such as autonomous driving and medical diagnosis [25], [43]. Many efforts have been made to interpret deep neural networks from different perspectives. Typically, researches include part-based methods [45],[47], attributes-based methods [14],[13], saliency maps [1], [12], [46], activation maximization [23], [24],

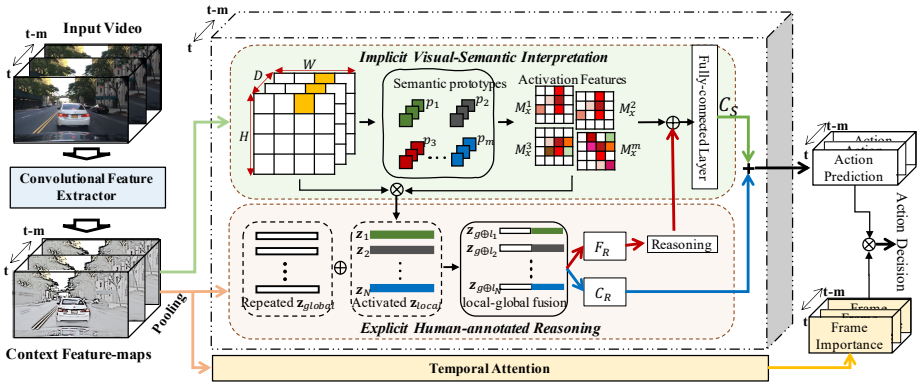

Fig. 1. Illustration of the proposed framework.

deconvolution/upconvolution to explain layers [10], [42] and have achieved inspiring progress to create human-interpretable black box models. However, such post-hoc solutions have limited capability in enhancing transparency and interpretability. Alternatively, prototype-based frameworks are proposed to build an inherently explainable architecture [2], [29], [28], [16], [22].

For the paradigm proposed by OIA, we argue that a human-defined finite explanation set provides inconsistent and insufficient explanations due to a lack of direct reasoning and failing to leverage temporal knowledge. In this work, we propose a novel prototype-based interpretable action decision making model (**InAction**) from implicit visual-semantic and explicit human-annotation perspectives. Different from prior prototype-based object recognition, our proposed model leverages and integrates action-inducing visual-semantic regions discovery, spatial relationships among objects, and temporal knowledge for driving decision prediction simultaneously and generates enriched explanations.

3 The Proposed Framework

3.1 Motivation

For autonomous driving, beyond pursuing high performance, interpretability is needed for safety-critical domains [25], [43]. This aims to imbue autonomous vehicles with reasoning abilities similar to human drivers. Existing efforts mainly adopt human-annotated explanations to guide system learning and generate human-understandable reasoning given the video inputs [39], which skews the model towards human annotation.

Unfortunately, human annotation has some drawbacks like insufficient explanation and inconsistent reasoning. Insufficient explanation means there are always implicit visual semantics not annotated by finite human-defined explanation set, which cannot be easily tracked through an end-to-end system with visual

inputs and explanation outputs. Inconsistent reasoning is particularly challenging since different people have different explanations, especially for complicated scenarios, leading to biases and insufficiency of the ground-truth annotation.

Motivated by this, we explore both the implicit visual-semantic interpretation and explicit human annotation jointly, and propose the Interpretable Action decision making model (**InAction**), whose goal is to enhance transparency and interpretability for autonomous driving action decision making.

3.2 Framework Architecture

An overview of the proposed InAction framework is shown as Figure 1. The model consists of a convolutional backbone $G(\cdot)$, and two interpretable action prediction modules—an implicit visual-semantic module and an explicit human-annotated reasoning module—to predict driving action and reasoning of the decision from different perspectives. Specifically, the implicit visual-semantic module is denoted as $G_S(\cdot)$, which takes the feature map per frame extracted by convolutional backbone as input to discover action-inducing concepts and the presence of learned semantic prototypes as visual cues for following prediction. For the explicit reasoning module, global visual features and the discovered action-inducing local regions are fused and input to two multi-task classifiers, predicting the driving action and human-annotated explanations, denoted as $C_R(\cdot)$ and $F_R(\cdot)$, respectively. Finally, the learned prototypical visual cues and predicted human-annotated explanations are fused and input to a fully-connected layer without bias as the action predictor, denoted as $C_S(\cdot)$. For the input video sequence, such prediction is applied to each frame, with a temporal attention layer employed to explore the contribution of each frame.

Mathematically, given an input video with m frames, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{m}$, whose action label as $\mathbf{y}_a \in \mathbf{A}$ and human annotated explanation $\mathbf{y}_e \in \mathbf{E}$, where $C_{\text{act}} = |\mathbf{A}|$ and $C_{\text{exp}} = |\mathbf{E}|$ are the numbers of categories of actions and human-annotated explanations, respectively. For each frame $\mathbf{x}$, the convolutional backbone extracts the feature map $\mathbf{f} = G(\mathbf{x})$ with shape $H \times W \times D$, where W and H denote the width and height, respectively, and D is the number of channels. For the clarity of description, denoting all the patches in the feature map as $\mathbf{Z_x} = \{\mathbf{z}_i \in \mathbf{f}\}_{i=1}^{HW}$, and the shape of each patch $\mathbf{z}_i$ is $\mathbb{R}^{D\times 1\times 1}$. The implicit visual-semantic module will slide over the whole feature map and calculate the activation scores for all patches in the feature map with respect to the presence of learned semantic prototypes. On the one hand, those regions primarily activated corresponding to specific prototypes are selected as action-inducing semantic regions and being fused with the global features to predict the action and explicit human-annotated explanation. On the other hand, the limitations of the activation map will be compensated by the predicted human-annotated explanations for the action prediction.

Implicit Visual Semantic Interpretation

To explore the action-inducing local regions in the visual input, we assign m_k semantic prototypes for each action class k, resulting in $m = m_k \times C_{\text{act}}$ prototypes in total, making up the visual-semantic layer $\mathbf{P} = \{\mathbf{P}_k\}|_{k=1}^{C_{\text{act}}}$, in which

$\mathbf{P}_k = \{\mathbf{p}_j\}|_{j=1}^{m_k}$, and $\mathbf{p}_j$ denotes the semantic visual prototypes to be learned for predicting action class k. Given the convolutional output feature map $\mathbf{Z_x}$ and prototype $\mathbf{p}_j$, the visual-semantic layer will go though all patches $\mathbf{z}_i \in \mathbf{Z_x}$ of the feature map to compute the activation score between them:

$$s_{ij} = \log\Big(\frac{\|\mathbf{z}_i - \mathbf{p}_j\|^2 + 1}{\|\mathbf{z}_i - \mathbf{p}_j\|^2 + \epsilon}\Big), \tag{1}$$

where ϵ is a small positive value, and the activation score s_{ij} represents how strongly a semantic prototype is presented in the specific region of the input frame. The activation scores of all the patches in the feature map produce an activation heat map $\mathbf{M}_\mathbf{x}^j$ with shape $H \times W$, identifying how similar each part of the input frame is to one specific prototype $\mathbf{p}_j$. Calculating activation maps for all prototypes results in an activation feature set $\mathbf{M_x} = \{\mathbf{M}_\mathbf{x}^j\}_{j=1}^m, \mathbf{M}_\mathbf{x}^j \in \mathbb{R}^{H\times W}$.

Intuitively, the most important patches for making action decision should be clustered around semantically similar prototypes of each specific action category, and the clusters centered at prototypes from different action categories are well separated. Thus, we also adopt a discriminative prototype learning loss as:

$$\mathcal{L}_d = \lambda_1 \mathbb{E}_{\mathbf{x}\in\mathbf{X}} \min_{\mathbf{p}_j \in \mathbf{P}_{\mathbf{y}_a}} \min_{\mathbf{z}\in\mathbf{Z_x}} \|\mathbf{z} - \mathbf{p}_j\|^2 - \lambda_2 \mathbb{E}_{\mathbf{x}\in\mathbf{X}} \min_{\mathbf{p}_j \notin \mathbf{P}_{\mathbf{y}_a}} \min_{\mathbf{z}\in\mathbf{Z_x}} \|\mathbf{z} - \mathbf{p}_j\|^2, \tag{2}$$

where λ_1 and λ_2 are two hyper-parameters determining the contributions of the two loss terms. Minimizing $\mathcal{L}_d$ encourages that every input frame at least has one prototype from its own action strongly activated in one of its latent feature map patches, while maximizing the distances between the patches and the prototypes from different classes. Such an optimization objective shapes the latent space into a semantically meaningful clustering structure.

Explicit Human-annotated Reasoning

Compared to implicit region-based action-inducing prototypes searching, human-annotated reasoning explains the driving decision in a more intuitive and abstract way. Normally natural language annotation involves temporal and spatial knowledge from visual inputs, which provides a more high-level explanation to the decision making. Intuitively, such explanation includes the global scene understanding and corresponding action-inducing objects.

Inspired by OIA [39], we propose an Explicit Human-annotated Reasoning module in a multi-task fashion to jointly generate human-annotated explanations and predict action. Specifically, for all the patches in the extracted feature map, we select top-N patches that activate any one of the prototypes assigned to the same action class as the action-inducing local components, denoted as $\mathbf{Z}_{local} = \{\mathbf{z}_l\}_{l=1}^N$, where $\mathbf{z}_l \in \mathbf{Z_x}$. The activation scores denote the importance of such patches contributing to the action decision making. It is noteworthy that the action-inducing local components $\mathbf{Z}_{local}$ are the presence of specific learned semantic prototypes, thus are not limited to be objects detected by the pre-trained object detection backbone, which is one of the limitations of OIA [39]. The selected top-N most activated patches can represent various scene contexts, environmental information, in addition to human-defined objects. Furthermore,

we consider that the global feature map provides an overall understanding of the visual input and the information like environmental status, e.g., "Road is clear", and agent relationship, e.g., "There is a vehicle parking on the right". In this sense, the local action-inducing components are concatenated with the global features, then input into to the action predictor $C_R(\cdot)$ and human-annotated explanation predictor $F_R(\cdot)$.

Specifically, the global feature map $\mathbf{Z_x}$ is processed with global average pooling and represented as a feature vector with the same dimension as each local patch $\mathbf{z}_l$, denoted as $\mathbf{z}_{global}$. Every local patch $\mathbf{z}_l$ is concatenated with the global feature $\mathbf{z}_{global}$ producing the local-global fused feature $\mathbf{Z}_{g\oplus l} = \{\mathbf{z}_l \oplus \mathbf{z}_{global}\}_{l=1}^{N}$, where $\mathbf{z}_l \in \mathbf{Z}_{local}$, and $\oplus$ is concatenation operation. The local-global feature is further vectorized then input to the following action and explanation prediction networks, optimizing the important local components that are highly associated with both action and explanation prediction. Eventually the predicted action and explanation are denoted as $\hat{\mathbf{y}}_a^R$ and $\hat{\mathbf{y}}_e^R$, respectively.

Considering the possible action decisions, we can explore to make a prediction with only one action or more than one action. If more than one action can be made, which is for a multi-label prediction task, the prediction logits are normalized by sigmoid function to the range between 0 and 1. If only one action can be made, which is a multi-class single-label task, the prediction logits are normalized by softmax function. Therefore, we formulate the multi-task learning objective of the explicit reasoning module as:

$$\mathcal{L}_r = L(\mathbf{y}_a, \hat{\mathbf{y}}_a^R) + L(\mathbf{y}_e, \hat{\mathbf{y}}_e^R), \tag{3}$$

where $L(\cdot, \cdot)$ denotes the cross-entropy loss and binary cross-entropy loss for single-label and multi-label prediction tasks, respectively.

Interpretable Decision Prediction So far, we design two kinds of explanations, i.e., $\mathbf{M_x}$ and $\hat{\mathbf{y}}_e^R$, for the decision making from two different perspectives. In order for these two explanations to interact and compensate for one another, the concatenated explanation vector $\hat{\mathbf{y}}_e = [\mathbf{M_x}, \hat{\mathbf{y}}_e^R]$ is exploited to a fully-connected layer $C_S(\cdot)$ to predict the action decision $\hat{\mathbf{y}}_a^S = C_S(\hat{\mathbf{y}}_e)$.

It is noteworthy that driver action decision making has more complicated scene contexts with many different agents, which is different from other prototype-based interpretable object recognition only considering the presence of some specific prototypical parts [2], [22], [29], [21]. Thus, the learned semantically meaningful prototypes that contribute to the final decision could be a part of or a complete object, even a set of objects or an environment region, in the input frame. Moreover, the location of a specific prototype, and the relationships between it with other objects and the environment, play crucial roles in determining the final action. Thus, rather than only choosing the maximum activation score for each prototype in the corresponding activation heat map, the whole activation feature set is considered for the fully-connected layer $C_S(\cdot)$ to integrate the spatial and relationship knowledge for predicting the action decision.

Similarly, we consider single-label and multi-label tasks with different activation functions and the learning objective of action prediction is defined as:

$$\mathcal{L}_s = L(\mathbf{y}_a, \hat{\mathbf{y}}_a^S), \tag{4}$$

where $L(\cdot, \cdot)$ represents cross-entropy loss for multi-class single-label tasks, while it is the binary cross-entropy loss for multi-label prediction tasks.

Cross-module Fusion and Temporal Aggregation Two action decision predictions $\hat{\mathbf{y}}_a^R$ and $\hat{\mathbf{y}}_a^S$ are obtained with different input knowledge. The former one is based on the visual features, while the latter one is based on explored explicit-and-implicit explanations. Thus, we accept two prediction logits followed by the specific activation function for multi-label or single-label problem, making the final aggregated action prediction, which is denoted as $\hat{\mathbf{y}}_a = \hat{\mathbf{y}}_a^R + \hat{\mathbf{y}}_a^S$.

Moreover, for the video input $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m$ with m frames, we make the decision prediction for each frame $\mathbf{x}_i \in \mathbf{X}$, resulting in a sequence of predictions $\{\hat{\mathbf{y}}_a^1, \ldots, \hat{\mathbf{y}}_a^m\}$. To find the most relevant information (key frames) in the observed sequence, a temporal attention layer is developed with a fully-connected layer followed by Softmax activation function, generating the importance δ_i for each frame $\mathbf{x}_i$. The objective with a temporal attention layer is defined as:

$$\mathcal{L}_t = L(\mathbf{y}_a, \sum\nolimits_{i=1}^{m} \delta_i \hat{\mathbf{y}}_a^i), \tag{5}$$

wherer $L(\cdot, \cdot)$ is cross-entropy loss or binary cross-entropy loss for single-label and multi-label prediction tasks, respectively.

Overall Objective. To sum up, we integrate two explanation modules into our unified framework and formulate the overall optimization objective as follows:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_r + \mathcal{L}_s + \mathcal{L}_t, \tag{6}$$

which includes two action decision classifiers and one explicit explanation predictor, and these two action decision classifiers will compensate for each other as they are based on different knowledge. In the test stage, we fuse the two predictions of action decision to obtain a more robust output.

4 Experiments

4.1 Experimental Setup

Pedestrian Situated Intent (PSI) dataset [4] contains 110 about 15 seconds long videos with 30 fps, and each is annotated with one of 3 speed change actions ("maintain speed", "slow down", and "stop") on frame level. The reasoning of the action decision is described in natural language, which will be used as explanation knowledge in our experiments. We split all videos into train/validation/test set with the ratio of 75%/5%/20%. We sample the tracks with length of 15 frames, and the overlap ratio is 0.8, while predicting the 16^{th} frame's action and explanation. Samples in PSI dataset are assigned one single label out of three

Table 1. Statistics of BDD-OIA and PSI dataset

Dataset	Action	# Frame	# Reasoning
BDD-OIA [39]	Forward	12,491	21 [Human-defined]
	Stop/Slow Down	10,432	
	Turn Left	5,902	
	Turn Right	6,541	
PSI [4]	Maintain Speed	5,800	29 [k-means clustered]
	Slow Down	4,925	
	Stop	1,177	

actions, so we evaluate the model by overall prediction accuracy and class-wise average accuracy for action prediction.

The original explanations are sentence-based, and each sentence contains descriptions of environmental context and human behaviors. We first split the original sentences into segments reflecting the environmental context or human behaviors. A syntactic dependency tree is applied to generate the dependency tagging of words, and then a set of heuristic rules are adopted to group each sentence into segments. Afterwards, the pre-trained BERT [8] is used to generate embeddings for all segments. The embedding of each segment is generated by averaging the embeddings of the words within the sentence segment. Consequently, we apply k-means clustering to obtain k semantic categories ($k = 29$ in our experiment). Given an explanation, since it is split into multiple segments and each might belong to different semantic categories, we generate k binary labels for each explanation to represent its semantics. For the human-annotated explanation, we report the overall F1 score and class-wise mean F1 score.

BDD-OIA dataset [39] is a subset of BDD100K [40] consisting of 22,924 5-second video clips, which were annotated with 4 action decisions ("move forward", "stop/slow down", "left turn", and "right turn") and 21 human-defined explanations. Specifically, each video contains at least 5 pedestrians or bicycle riders and more than 5 vehicles. The videos are collected with complex driving scenes to increase the scene diversity. Following the setting of [39], only the final frame of each video clip is used thus the temporal attention layer is neglected. As there are multiple possible action choices for each sample, we evaluate the performance by F1 score for each specific action, overall F1 score, and the class-wise average F1 score for both action and explanation prediction.

More statistics of the benchmarks are shown in Table 1.

Implementation Details. The Faster R-CNN [27] is pre-trained on the annotated images from BDD100K [40] and set as the backbone, which is followed by two 3×3 convolutional layers generating the global feature map with shape $7 \times 7 \times 256$ for each input frame. For implicit visual semantic interpretation module, we assign $m_k = 6$ prototypes with dimension 128 for each action class, resulting in $m = 24$ prototypes for BDD-OIA dataset, and $m = 18$ prototypes in total for PSI dataset. For our InAction model, we set $N = 10$ thus the $top - 10$ patches from the input feature map with the smallest distances compared to all

Table 2. Single-label action and multi-label explanation prediction on PSI dataset

Method	Maintain	Slow	Stop	act. Acc_{all}	act. mAcc	exp. $F1_{all}$	exp. mF1
OIA-global[39]	0.540	<u>0.774</u>	0.537	0.635	0.617	0.178	0.119
OIA [39]	0.693	0.622	0.463	0.643	0.593	0.189	0.110
Ours-f	<u>0.703</u>	0.771	<u>0.641</u>	<u>0.719</u>	<u>0.704</u>	<u>0.277</u>	<u>0.203</u>
Ours-v	**0.717**	**0.776**	**0.672**	**0.734**	**0.722**	**0.285**	**0.223**

Table 3. Multi-label action and explanation prediction on BDD-OIA dataset

Method	F	S	L	R	act. $F1_{all}$	act. mF1	exp. $F1_{all}$	exp. mF1
Res-101[39]	0.755	0.607	0.098	0.108	0.601	0.392	0.331	0.180
OIA[39]	**0.829**	**0.781**	**0.630**	**0.634**	**0.734**	**0.718**	0.422	0.208
OIA*[39]	0.792	0.742	0.594	<u>0.627</u>	0.705	0.689	0.501	0.293
Ours(proposals)	0.795	0.743	0.597	0.613	0.706	0.687	<u>0.558</u>	<u>0.332</u>
Ours(global)	<u>0.800</u>	<u>0.747</u>	<u>0.612</u>	0.619	<u>0.714</u>	<u>0.694</u>	**0.565**	**0.347**

semantic prototypes are selected to be fused with the global features for explicit human-annontated explanation and action prediction. The feature map is input to two additional 1×1 convolutional layers to reduce the channel dimension to be same as the prototypes dimension and normalized by sigmoid function following [2] before calculating the activation scores. The action predictor $C_S(\cdot)$ based on the fused explanation vector is one fully-connected layer without bias. We follow the same strategy of [2] to initialize and train the model. For the explicit human-annotated reasoning module, the action decision predictor $C_R(\cdot)$ is a three-layer fully-connected neural network, and the explanation predictor $F_R(\cdot)$ is two-layer fully-connected neural network. ReLU activation is used for all hidden layers. The model is optimized by Adam optimizer with learning rate initialized as 10^{-3}, and decayed by 0.1 every 10 epochs. For simplicity, we set $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ by default for all experiments. We empirically fix $m_k = 6$, and we observe the results are not sensitive to it if $m_k > 3$ on validation set.

4.2 Comparison Results

We compare our proposed InAction model with the OIA method [39] on the PSI and BDD-OIA datasets, and the results are reported in Table 2 and Table 3. OIA model only adopts the last frame of a sequence as input, thus we report two results produced by our model with only the last frame or the whole observed video sequence as input, denoted as Ours-f and Ours-v in Table 2, respectively. For experiments on BDD-OIA in Table 3, we reproduce the OIA model based on the official implementation released by the author, denoted as OIA*, in addition to the results reported by OIA [39]. The reproduced results of OIA on BDD-OIA are lower in action decision while better in explanation in term of F-1 score, compared with the reported OIA. Note that OIA adopts the detected

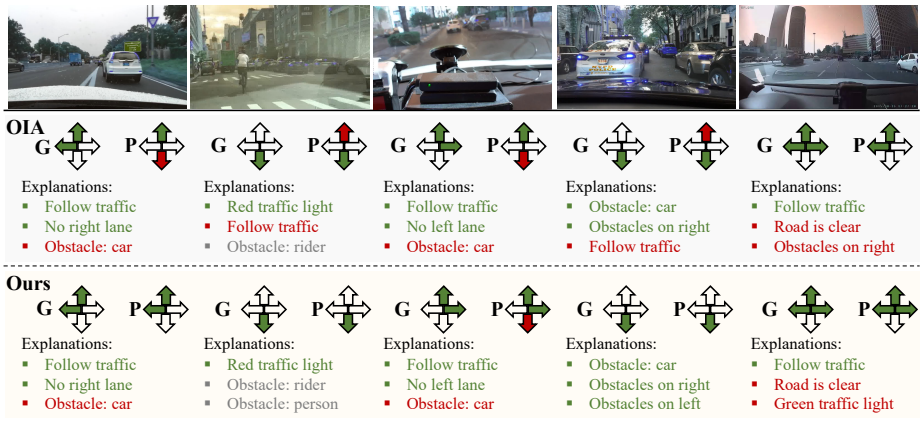

Fig. 2. Selected comparison examples of action and explicit explanation prediction between OIA and InAction on BDD-OIA dataset. **G** denotes the ground-truth annotation, and **P** shows the predicted result from OIA/Ours. green predictions are True Positive, red are False Positive, and gray are False Negative.

proposals generated by the backbone as local features. We utilize the implicit visual-semantic prototypes learned from the global feature map and from the detected proposals, and report the results as Ours(global) and Ours(proposals), respectively. Specifically, to obtain Ours(proposals), we extract the top-100 detected proposals features after average pooling process into the same size as the learned prototypes, then follow the same fusing strategy as aforementioned.

For the PSI results (Table 2), we notice that our proposed InAction model with only the last frame as input outperforms OIA around **0.07** and **0.01** for the overall and class-wise mean action prediction accuracy, respectively. When the whole video sequence is input to our model, the performance is improved further by 0.015 and 0.018, respectively, demonstrating our model can benefit from the temporal knowledge from the input sequence. The PSI dataset has an imbalanced distribution and there are much fewer samples belonging to the category "Stop", thus both OIA-global and OIA* obtain worse performance on this category compared to "Main speed" and "Slow down". Surprisingly, our model is able to achieve better performance on this decision. Moreover, as OIA adopts both global and local detection proposals as input for prediction, while InAction only uses the global feature map, so that we compare our model with another baseline OIA-global, which has the same architecture with OIA excluding the local proposal branch. From the results, we observe that OIA-global obtains worse overall performance compared to OIA and InAction.

From the BDD-OIA results (Table 3), we observe InAction can improve the action prediction performance compared to the reproduced OIA. For the reason prediction, we notice that the reproduced results outperform the numbers reported in the OIA paper around 0.08, and our proposed method can further improve the overall F1 and class-wise mean F1 both over **0.5**. This demonstrates

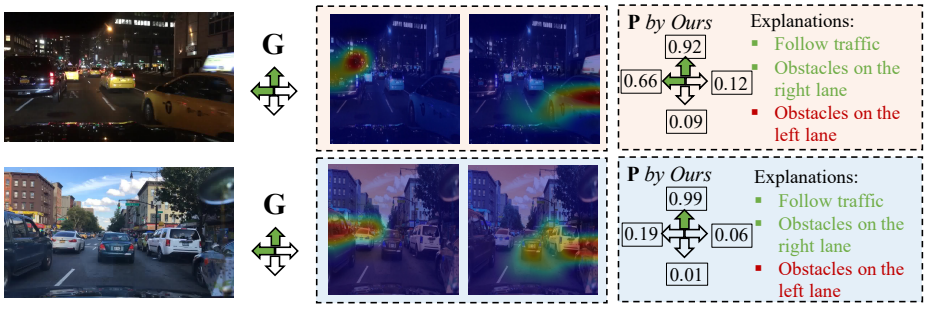

Fig. 3. Comparison of explanations produced by the implicit visual-semantic module and the explicit human-annotated reasoning module for examples on BDD-OIA.

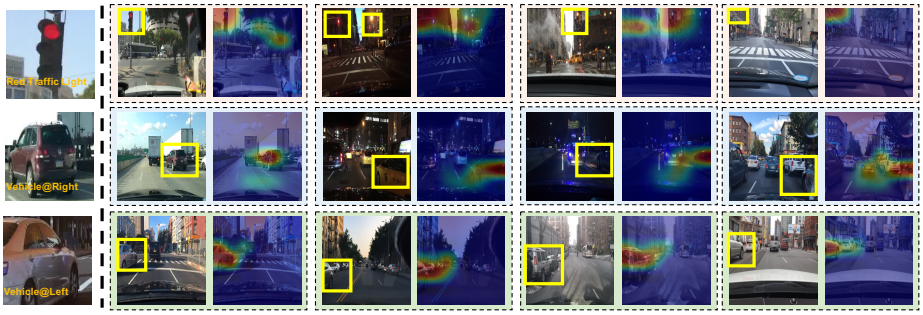

Fig. 4. Visualizing prototypes by selecting the most similar patches from the training samples, where each row shows one explanation.

that our model works well in both action prediction and explanation reasoning. Moreover, we observe that the results produced with prototypes learned on the global feature maps are better than based on the detected proposals. We argue that relying on the detected proposals will make the model fail, and constrain the representative capabilities of learned semantic prototypes, compared to exploring the implicit visual-semantic knowledge based on the whole input image.

4.3 Interpretability Analysis

Comparison with OIA. We present qualitative results in Figure 2 to demonstrate the interpretability and transparency of the propose InAction model. For the same visual input, we compare both action and explanation prediction of OIA and our InAction. From the selected examples, we notice that OIA made wrong action predictions while InAction can achieve correct results in some cases. The only wrong prediction in the 3^{rd} example is that both OIA and the explicit human-annotated reasoning module in InAction recognize the white vehicle in front and predict the explanation as "Obstacle: car", then make the "Stop/Slow down" decision. However, the ground-truth action annotation does not contain

this label. Such an observation demonstrates that insufficient explanation and inconsistency reasoning always exists in the human-defined annotations, especially on single-frame based prediction tasks.

Compensation between Implicit and Explicit Interpretation. In Figure 3, we compare the generated explanations from the implicit and explicit modules for the same task. We notice that some human-annotated reasoning are also captured by the implicit semantic prototypes, e.g., "Obstacles on the right lane". However, some explanations discovered by the implicit visual prototypes compensate the lack of human annotation. For example, the vehicle on the left lane in the second row example is quite close but not annotated, and the ground-truth label is "forward" and "turn left", while fortunately, our model notices the obstacle on the left lane and predict "forward" only.

Implicit Visual Semantics Analysis.

To illustrate the learned implicit visual semantic prototypes in an intuitive way, we visualize the prototypes via the most similar patches of images in the BDD-OIA dataset [2]. Figure 4 shows the selected examples with patches highly activated by specific semantic prototypes from action decision "Stop", "Turn left", and "Turn right". The most activated patch of the given input for selected prototypes are marked by bounding boxes in the original input, which represent the image patches that InAction considers to focus on corresponding to specific prototypes. From the results, we observe that when the implicit visual-semantic reasoning module slides over the whole input to obtain activation map, these three prototypes are represented as "Red traffic light", "Vehicle at right", and "Vehicle at left", respectively. Any region is strongly activated by one of the specific prototypes, or, in other words, one of the prototypes presents strongly in the input frame, will play a crucial role in the final prediction.

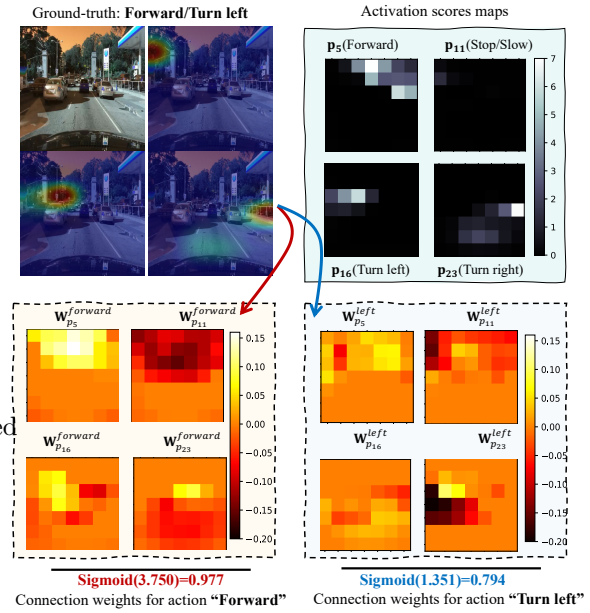

Fig. 5. Visualization of reasoning of selected instance.

Reasoning Process of InAction. Prior prototype-based models only observe the most strongly activated region. However, driving action prediction has much

more complicated scene context and multiple objects involved as hints, so the spatial location of each prototype presence and the relationships among different components make crucial influence for the final decision prediction. Figure 5 shows the reasoning process of our InAction predicting the action decision for a test sample, which is annotated as "Forward/Turn left". Given the input frame, the implicit visual semantic interpretation module compares every patch in the feature map against the learned prototypes, producing the activation score maps. The activation maps that are most strongly activated by prototypes are shown as the top-right heatmaps in Figure 5, where $\mathbf{p}_5, \mathbf{p}_{11}, \mathbf{p}_{16}, \mathbf{p}_{23}$ are assigned to action classes "Forward", "Stop/Slow down", "Turn left", and "Turn right", respectively. Although m prototypes are assigned to C action decision resulting in m_k prototypes per class during training, all activation scores produced by m prototypes over the feature map of the input frame are multiplied by the weight matrix in the last fully-connected layer $C_S(\cdot)$ to generate the output prediction. The weights in the fully-connected layer represent the connections between prototypes and the predicted classes. In Figure 5, we select the weights ($\mathbf{W}$) for class "Forward" and "Turn left" corresponding to the selected prototypes, and show them after reshaped into the same shape as the activation map. From the weights over different regions of the feature map/activation map, we observe that the same prototype plays different roles for different action decisions. For example, components similar to prototype $\mathbf{p}_{11}$ appearing in the top area of the view will make negative contribution to the prediction of "Forward", while for the prediction of class "Turn left", it will reduce the probability of "Turn left" only when it appears at the top-left corner, otherwise, this prototype is comparably neutral. Interestingly, the prototype shown in the first row of Figure 4 is prototype $\mathbf{p}_{11}$, which represents "Red traffic light".

5 Conclusion

In this paper, we developed a novel Interpretable Action (**InAction**) decision making model to provide enriched explanations from both explicit human annotation and implicit visual semantics perspectives. To implement this, two interpretable modules were proposed including a visual semantic module and an explicit reasoning module. Specifically, the first module aimed to capture the region-based action-inducing semantic concepts from the visual inputs, so that our model could automatically learn the implicit visual cues to provide a human-understandable explanation. The second module attempted to benefit from the human-annotated reasoning for action decision making so that our model was able to provide a more high-level interpretation by aligning visual inputs to human annotations. Experimental results on two autonomous driving benchmarks demonstrated the effectiveness of our **InAction** model.

Acknowledgment. We thank the Toyota Collaborative Safety Research Center for funding support.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
2. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.: This looks like that: Deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems. pp. 8928–8939 (2019)
3. Chen, L., Yang, T., Zhang, X., Zhang, W., Sun, J.: Points as queries: Weakly semi-supervised object detection by points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8823–8832 (June 2021)
4. Chen, T., Tian, R., Chen, Y., Domeyer, J., Toyoda, H., Sherony, R., Jing, T., Ding, Z.: Psi: A pedestrian behavior dataset for socially intelligent autonomous car. arXiv preprint arXiv:2112.02604 (2021)
5. Choi, C., Choi, J.H., Li, J., Malla, S.: Shared cross-modal trajectory prediction for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 244–253 (June 2021)
6. Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1601–1610 (June 2021)
7. Darms, M., Rybski, P., Urmson, C.: Classification and tracking of dynamic objects with multiple sensors for autonomous driving in urban environments. In: 2008 IEEE Intelligent Vehicles Symposium. pp. 1197–1202 (2008). https://doi.org/10.1109/IVS.2008.4621259
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019)
9. Dong, J., Chen, S., Zong, S., Chen, T., Labi, S.: Image transformer for explainable autonomous driving system. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 2732–2737 (2021)
10. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4829–4837 (2016)
11. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems (2020)
12. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision. pp. 3429–3437 (2017)
13. Jing, T., Liu, H., Ding, Z.: Towards novel target discovery through open-set domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9322–9331 (2021)
14. Jing, T., Xia, H., Hamm, J., Ding, Z.: Augmented multi-modality fusion for generalized zero-shot sketch-based visual retrieval. IEEE Transactions on Image Processing (2022)

15. Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5830–5840 (June 2021)
16. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15719–15728 (2021)
17. Kim, J., Canny, J.: Interpretable learning for self-driving cars by visualizing causal attention. In: Proceedings of the IEEE international conference on computer vision. pp. 2942–2950 (2017)
18. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–578 (2018)
19. Li, J., Zhan, W., Hu, Y., Tomizuka, M.: Generic tracking and probabilistic prediction framework and its application in autonomous driving. IEEE Transactions on Intelligent Transportation Systems **21**, 3634–3649 (2020)
20. Li, P., Qin, T., Shen, a.: Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
21. Ming, Y., Xu, P., Qu, H., Ren, L.: Interpretable and steerable sequence learning via prototypes. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 903–913 (2019)
22. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14933–14943 (2021)
23. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in neural information processing systems **29**, 3387–3395 (2016)
24. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2**(11), e7 (2017)
25. Omeiza, D., Webb, H., Jirotka, M., Kunze, L.: Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems pp. 1–21 (2021)
26. Pang, B., Zhao, T., Xie, X., Wu, Y.N.: Trajectory prediction with latent belief energy-based model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11814–11824 (June 2021)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
28. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
29. Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B.: Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1420–1430 (2021)
30. Shafiee, N., Padir, T., Elhamifar, E.: Introvert: Human trajectory prediction via conditional 3d attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16815–16825 (June 2021)
31. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8994–9003 (June 2021)

32. Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., Zhang, H.: A comparative study of real-time semantic segmentation for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
33. Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Muhammad, N.: A survey of end-to-end driving: Architectures and training methods. IEEE Transactions on Neural Networks and Learning Systems (2020)
34. Wang, D., Devin, C., Cai, Q.Z., Krähenbühl, P., Darrell, T.: Monocular plan view networks for autonomous driving. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2876–2883. IEEE (2019)
35. Wang, D., Devin, C., Cai, Q.Z., Yu, F., Darrell, T.: Deep object-centric policies for autonomous driving. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8853–8859. IEEE (2019)
36. Xia, H., Ding, Z.: Hgnet: Hybrid generative network for zero-shot domain adaptation. In: European Conference on Computer Vision. pp. 55–70. Springer (2020)
37. Xia, H., Ding, Z.: Structure preserving generative cross-domain learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4364–4373 (2020)
38. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2174–2182 (2017)
39. Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y., Vasconcelos, N.: Explainable object-induced action decision for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9523–9532 (2020)
40. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
41. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access **8**, 58443–58469 (2020)
42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
43. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence (2021)
44. Zhang, Z., Fidler, S., Urtasun, R.: Instance-level segmentation for autonomous driving with deep densely connected mrfs. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 669–677 (2016)
45. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 5209–5217 (2017)
46. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
47. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–134 (2018)