

OIL: Observational Imitation Learning

<https://sites.google.com/kaust.edu.sa/oil/>

Guohao Li*, Matthias Müller*, Vincent Casser, Neil Smith, Dominik L. Michels, Bernard Ghanem
 Visual Computing Center, KAUST, Thuwal, Saudi Arabia

{guohao.li, matthias.mueller.2, vincent.casser, neil.smith, dominik.michels, bernard.ghanem}@kaust.edu.sa

Abstract—Recent work has explored the problem of autonomous navigation by imitating a teacher and learning an end-to-end policy, which directly predicts controls from raw images. However, these approaches tend to be sensitive to mistakes by the teacher and do not scale well to other environments or vehicles. To this end, we propose Observational Imitation Learning (OIL), a novel imitation learning variant that supports online training and automatic selection of optimal behavior by observing multiple imperfect teachers. We apply our proposed methodology to the challenging problems of autonomous driving and UAV racing. For both tasks, we utilize the Sim4CV simulator [23] that enables the generation of large amounts of synthetic training data and also allows for online learning and evaluation. We train a perception network to predict waypoints from raw image data and use OIL to train another network to predict controls from these waypoints. Extensive experiments demonstrate that our trained network outperforms its teachers, conventional imitation learning (IL) and reinforcement learning (RL) baselines and even humans in simulation.

I. INTRODUCTION

In the machine learning community, solving complex sequential prediction problems usually follows one of two different paradigms: reinforcement learning (RL) or supervised learning (SL), more specifically imitation learning (IL). On the one hand, the learner in conventional IL is required to trust and replicate authoritative behaviors of a teacher. The drawbacks are primarily the need for extensive manually collected training data and the inherent subjectivity to potential negative behaviors of teachers, since in many realistic scenarios they are imperfect. On the other hand, RL does not specifically require supervision by a teacher, as it searches for an optimal policy that leads to the highest eventual reward. However, a good reward function, which offers the agent the opportunity to learn desirable behaviors, requires tedious and meticulous reward shaping [26]. Recent methods have used RL to learn simpler tasks without supervision [6], but they require excessive training time and a very fast simulation (e.g. 1 000 fps). In this paper, we demonstrate that state-of-the art performance can be achieved by incorporating RL concepts into direct imitation to learn only the successful actions of multiple teachers. We call this approach Observational Imitation Learning (OIL). Unlike conventional IL, OIL enables learning from multiple teachers with a method for discarding bad maneuvers by using a reward based online evaluation of the teachers at training time. Furthermore, our approach allows for a modular architecture that abstracts perception and control, which allows for more flexibility when training in diverse environments with different visual properties and control dynamics. In our experiments, it



Fig. 1: OIL-trained Autonomous Driving (*left*) and OIL-trained UAV Racing (*right*) on test tracks created using Sim4CV [23].

is shown that this approach leads to greater robustness and improved performance, as compared to various state-of-the-art IL and RL methods. Moreover, OIL allows for control networks to be trained in a fully automatic fashion requiring no human annotation but rather can be trained using automated agents. We demonstrate that our approach outperforms other state-of-the-art end-to-end network architectures and purely IL and RL based approaches.

We apply OIL to both autonomous driving and UAV racing in order to demonstrate the diverse scenarios in which it can be applied to solve sequential prediction problems. We follow recent work [6] that tests AI systems through the use of computer games. We use Sim4CV based on the Unreal Engine 4 (UE4), which has both a residential driving environment with a physics based car and gated racing tracks for UAV racing [23]. The simulator is multi-purpose, as it enables the generation of synthetic image data, reinforcement based training in real-time, and evaluation on unseen tracks. We demonstrate that using OIL enables to train a modular neural network predicting controls for autonomous driving and the more complex task of UAV racing in the simulated Sim4CV environment. Through extensive experiments, we show that OIL outperforms its teachers, conventional IL and RL approaches and even humans in simulation.

Contributions. (1) We propose Observational Imitation Learning (OIL) as a new approach for training a stationary deterministic policy that overcomes shortcomings of conventional imitation learning by incorporating reinforcement learning ideas. It learns from an ensemble of imperfect teachers, but only updates the policy with the best maneuvers of each teacher, eventually outperforming all of them. (2) We use a flexible network architecture which adapts well to different perception and control scenarios. We show that it is suitable for solving complex navigation tasks (e.g. autonomous driving and UAV racing). (3) To the best of our knowledge, this paper is the first to apply imitation learning to multiple teachers while being robust to teachers that exhibit bad behavior.

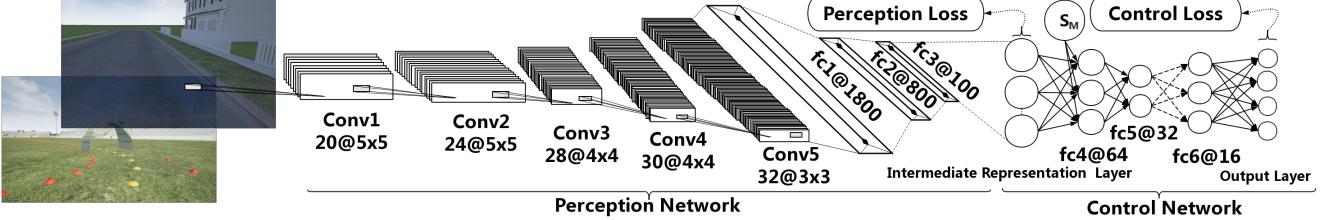


Fig. 2: Pipeline of our modular network for autonomous navigation. The perception network ϕ takes the raw image as input and predicts waypoints. The control network φ takes waypoints and vehicle state as input and outputs an appropriate control signal, e.g. throttle (T), aileron (A), elevator (E), and rudder (R) for the UAV and only gas (E) and steering (R) for the car.

II. RELATED WORK

The task of training an actor (e.g. ground vehicle, human, or UAV) to physically navigate through an unknown environment has traditionally been approached either through supervised learning (SL) and in particular imitation learning (IL), reinforcement learning (RL), or a combination of the two. A key challenge is learning a high dimensional representation of raw sensory input within a complex 3D environment. Similar to our approach, many recent works such as [8, 30, 31, 39, 15, 2] use modern game engines or realistic physics-based simulators to evaluate complex navigation tasks.

Imitation Learning (IL). In physics based navigation, IL can be advantageous when high dimensional feedback can be recorded. It has been applied to both autonomous driving [3, 5, 40, 47, 27] and UAV navigation [40, 13, 22, 2, 41]. In particular, DAGGER [34] and its variants [2] [27] have been widely used for many robotic control tasks. However, limitations such as requirements of significant fine-tuning, inability to predict human-like controls, data augmentation being only corrective and the requirement for near-optimal expert data makes DAGGER hard to scale up to more complex problems. Moreover, the flaws and mistakes of the teachers are learned along with differing responses. See Section V-B for the comparison of two popular IL methods (Behavioral Cloning [45] and DAGGER) to OIL. Follow-up work such as AGGREGATE [33] and Deeply AggreVaTeD [42] attempt to mitigate these problems by introducing exploratory actions and measuring actual induced cost instead of optimizing for expert imitation only. They also claim exponentially higher sample efficiency than many classic RL methods. A number of improvements in other respects have been published, such as SafeDAgger [48] that aims to make DAGGER more (policy) query-efficient and LOLS [4] that aims to improve upon cases where the reference policy is sub-optimal.

Reinforcement Learning (RL). RL provides an alternative to IL by using rewards and many iterations of exploration to help discover the proper response through interactive trial and error. Recent work on autonomous driving has employed RL [21, 19, 14, 36, 35, 12, 7]. RL networks may not be able to discover the optimal outputs in higher-order control tasks. For example, Dosovitskiy et al. [7] find RL to under-perform in vehicle navigation due to the extensive hyperparameter space.

RL methods can be divided into three classes: value-based, policy-based, and actor-critic based methods [43]. In particular, actor-critic based methods, e.g. A3C [21] and DDPG [19], are notably the most popular algorithms in the RL community. However, achieving strong results with RL is difficult, since it is very sensitive to the reward function, it can be sample inefficient, and it requires extensive training time due to the large policy space (see Section V-B for comparison of DDPG to OIL). Methods such as TRPO [38] have been developed to provide monotonic policy improvements in most cases, but still require extensive training time.

Combined Approaches. Several methods exist that combine the advantages of IL and RL. Most of them focus on tackling the problem of low RL sample efficiency by providing suitable expert demonstrations ([18, 49, 44, 29, 9, 17]) and guided policy search ([16, 6, 10, 49]). Others focus on risk awareness as real-world deployment failures can be costly [1]. Generative adversarial imitation learning [11] avoids the costly expense of IRL by directly learning a policy from supplied data. It explores randomly to determine which actions bring a policy closer to an expert but never directly interacts with the expert as in DAGGER. The authors note a combination of both random search and interaction with the expert would lead to better performance, an insight employed in our approach. Similar to our approach, [10] apply Bayesian Model Combination to guide the policy to learn the best combination of multiple experts. The recent CFN [25] algorithm is the most related method to our approach; trajectories of multiple controllers are filtered and then fused by a neural network in order to obtain a more robust controller.

We draw inspiration for OIL from these hybrid approaches. In contrast to pure IL, OIL can prevent itself from learning bad demonstrations from imperfect teachers by observing teachers' behaviours and estimating the advantage or disadvantage to imitate them. Unlike pure RL, it converges to a high performance policy without too much exploration since it is guided by the best teacher behaviors. While sharing the advantage of higher sample efficiency with other hybrid approaches, our method has the specific advantage of inherently dealing well with bad demonstrations, which is a common occurrence in real-world applications.

III. METHODOLOGY

After providing a brief review of related learning strategies for sequential decision making (i.e. Markov Decision Process, Imitation Learning, and Reinforcement Learning), we introduce our proposed Observational Imitation Learning (OIL), which enables automatic selection of the best teacher (among multiple teachers) at each time step of online learning.

A. Markov Decision Process

OIL is a method that enables a learner to learn from multiple sub-optimal or imperfect teachers and eventually to outperform all of them. To achieve this goal, training needs to be done by repeatedly interacting with an environment \mathcal{E} . We consider the problem as a Markov Decision Process (MDP) consisting of an agent and environment. At every time step n , the agent observes a state s_n or a partial observation o_n of state s_n . In our setting, we assume the environment is partially-observed but we use s_n to represent the state or the partial observation for simplicity. Given s_n , the agent performs an action a_n within the available action space \mathcal{A} based on its current policy π , where θ represents the parameters of the policy. Then, the environment \mathcal{E} provides the agent a scalar reward r_n according to a reward function $R(s_n, a_n)$ and transfers to a new state s_{n+1} within the state space \mathcal{S} under the environment transition distribution $p(s_{n+1}|s_n, a_n)$. After receiving an initial state s_1 , the agent generates a trajectory $\tau = (s_1, a_1, s_2, a_2, \dots, s_N)$ after N time steps. The trajectory can end after a certain number of time steps or after the agent reaches a goal or terminal state. The objective of solving a MDP problem is to find an optimal policy π^* from policy space Π that maximizes the expected sum of discounted future rewards, R_n at time n .

It is commonly referred to as the value function:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}[R_n | s_n = s, \pi] \\ &= \mathbb{E}\left[\sum_{n'=n}^N \gamma^{n'-n} r_{n'} | s_n = s, \pi\right], \end{aligned} \quad (1)$$

where $\gamma \in [0, 1]$ is a discounted factor that trades off the importance of immediate and future rewards, and N is the time step when the trajectory τ terminates. The optimal policy π^* maximizes the value function for all $s \in \mathcal{S}$:

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi} V^\pi(s) \quad (2)$$

B. Imitation Learning

Imitation learning (IL) is a supervised learning approach to solve sequential decision making problems by mimicking an expert policy π^e . Instead of directly optimizing the value function, IL minimizes a surrogate loss function $\mathbb{L}(s, \pi, \pi^e)$. Let d_π denote the average distribution of visited observations when an arbitrary policy π is executed for T time steps. Behavioral Cloning, one of the simplest IL algorithms, trains a learner policy network π^l to fit the input (observations) and output (actions) of the expert policy by minimizing the surrogate loss [34]:

$$\pi^l = \operatorname{argmin}_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\pi^e}} [\mathbb{L}(s, \pi, \pi^e)] \quad (3)$$

However, this leads to poor performance because the encountered observation spaces of the learner and the expert are different, thus violating the independent, identically distributed (i.i.d.) assumption of statistical learning approaches and causing compounding errors [32]. DAGGER (Dataset Aggregation) [34] alleviates these errors in an iterative fashion by collecting the state-action pairs visited by the learned policy, but labeled by the expert. Its goal is to find a policy π^l that minimizes the surrogate loss under the observation distribution d_π induced by the current policy π :

$$\pi^l = \operatorname{argmin}_{\pi \in \Pi} \mathbb{E}_{s \sim d_\pi} [\mathbb{L}(s, \pi, \pi^e)] \quad (4)$$

A major drawback of DAGGER is that it highly depends on the expert's performance, while a near-optimal expert is hard to acquire for most tasks.

C. Reinforcement Learning

Reinforcement Learning (RL) is another category of methods to solve the MDP problem by trial and error. RL methods can be divided into three classes: value-based, policy-based, and actor-critic-based methods [43]. Specifically, actor-critic-based methods, e.g. A3C [21] and DDPG [19], are currently the most popular algorithms in the RL community. The most related actor-critic method to our work is the Advantage Actor-critic approach [21], which uses an advantage function for policy updates instead of the typical value function or Q-function [20]. Intuitively, this advantage function evaluates how much improvement is obtained if action a is taken at state s as compared to the expected value. It is formally defined as follows [28]:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s) \quad (5)$$

D. Observational Imitation Learning (OIL)

As discussed in Section III-B, imitation learning requires a near-optimal teacher and extensive augmentation for exploration. Getting labeled expert data is expensive and not scalable. While RL approaches (Section III-C) do not require supervision and can freely explore an environment, they are time-consuming and may never learn a good policy unless the reward is very well designed. In an effort to combine the strengths of both approaches we propose *Observational Imitation Learning* (OIL). Inspired by advantage actor-critic learning [21], we learn to imitate only the best behaviours of several sub-optimal teachers. We do so by estimating the value function of each teacher and only keeping the best to imitate. As a result, we can learn a policy that outperforms each of its teachers. Learning from multiple teachers also allows for exploration, but only of feasible states, leading to faster learning than in RL.

Since we do not require expert teachers we can obtain labeled data much more cheaply. We assume easy access to K cheap sub-optimal teacher policies (e.g. simple PID controllers). We denote the teacher policy set as $\Pi^t = \{\pi^{t_1}, \pi^{t_2}, \dots, \pi^{t_K}\}$, where π^{t_k} is the teacher policy corresponding to teacher k and denote the learner policy as π^l .

The advantage function of current learner policy π^l compared to teacher policy π^t at state s can be written as:

$$A(s, \pi^l, \pi^t) = V^{\pi^l}(s) - V^{\pi^t}(s) \quad (6)$$

The advantage function $A(s, \pi^l, \pi^t)$ determines the advantage of taking the learner policy π^l at state s compared to the teacher policy π^t where π^l and π^t are analogous to an actor and critic policy respectively. Note that the advantage function in regular RL settings is used to cope with the learner policy. In contrast, our advantage function Equation 6 considers both the learner policy and the teacher policy. In multi-teacher scenarios, we select the most critical teacher π^{t*} as the critic policy (refer to Algorithm 1).

We define our training in terms of observation rounds, each consisting of three phases: *observing, rehearsing, acting*.

Observe. We estimate the value functions $V^\pi(s)$ of the learner policy as well as of all teacher policies using Monte-Carlo sampling (rolling out the trajectory of each policy to get the returns R_t). We select the most critical teacher policy $\pi^{t*} = \operatorname{argmax}_{\pi^t \in \Pi^t} V^{\pi^t}(s)$ as the critic. Then we compute

the advantage function $A(s_1, \pi^l, \pi^{t*})$ between the learner policy and the most critical teacher policy. If the advantage $A(s_1, \pi^l, \pi^{t*}) < 0$ (i.e. there exists a sub-optimal teacher policy with a higher advantage than the current learner policy), we enter the *rehearsing* phase. Otherwise, we go to the *acting* phase directly.

Rehearse. After computing $A(s_1, \pi^l, \pi^{t*})$, we can optimize the policy by actor-critic methods or by optimizing the surrogate loss. In order to benefit from the fast convergence of IL instead of optimizing the advantage function directly, we optimize the surrogate loss in Equation 7 iteratively as follows:

$$\pi^{l(i+1)} = \operatorname{argmin}_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\pi^{l(i)}}} [\mathbb{L}(s, \pi^{l(i)}, \pi^{t*})] \quad (7)$$

where $\pi^{l(i)}$ is the learner policy at the i -th iteration. In our implementation, we use a DNN to represent our learner policy as $\pi^l(s|\theta)$, where θ represents the parameters of the neural network. In order to minimize the surrogate loss, we roll out the learner (actor) policy and use the selected teacher (critic) to correct the learner's actions. In other words, we minimize the surrogate loss with states encountered by the learner policy and actions labeled by the most critical teacher policy. We minimize the surrogate loss by performing gradient descent with respect to θ on collected data \mathcal{D} and update the learner policy until $A(s_1, \pi^l, \pi^{t*}) > \epsilon$ or I episodes. In practice, the advantage function A can be estimated using *Monte Carlo Methods* or *Bootstrapping Methods*. In our experiment, we use *Monte Carlo Methods* to estimate the advantage function by rolling out policies (see Section V-A).

Act. After rehearsing, the learner policy will perform well from the initial state s_1 ; we roll out the current policy π^l to the new state s'_1 by acting J steps.

IV. NETWORK AND TRAINING DETAILS

In this section we present a modular network architecture based on CFN [25] for autonomous driving and UAV racing

Algorithm 1: Observational Imitation Learning (OIL).

```

Initialize Learner training database  $\mathcal{D} \leftarrow \emptyset$ ;
Initialize Learner network  $\pi^l(s|\theta)$  with random weights  $\theta$ ;
for observation round  $m \leftarrow 1$  to  $M$  do
    Receive initial state  $s_1$  from the environment;
    Estimate learner value function  $V^{\pi^l}(s_1)$  ;
    Estimate teacher value functions  $V^{\pi^t}(s_1)$ ,  $\forall \pi^t \in \Pi^t$ ;
    Choose  $\pi^{t*} = \operatorname{argmax}_{\pi^t \in \Pi^t} V^{\pi^t}(s_1)$ ;
    Compute advantage function  $A(s_1, \pi^l, \pi^{t*})$ ;
    while  $A(s_1, \pi^l, \pi^{t*}) < 0$  do
        repeat
            Sample N-step trajectories using learner policy;
            while  $n < N$  do
                repeat
                     $a_n \sim \pi^l(s_n|\theta)$ ;
                    Take action  $a_n$ , observe  $r_n$   $s_{n+1}$  ;
                    Add state-action  $(s_n, \pi^{t*}(s_n))$  to  $\mathcal{D}$ ;
                    Update  $\pi^l(s|\theta)$  by minimizing
                     $\mathbb{L}(s, \pi^l, \pi^{t*})$  from  $\mathcal{D}$ ;
                     $s_n \leftarrow s_{n+1}$ ;
                until  $s_n$  is a terminal state;
            end
        until  $A(s_1, \pi^l, \pi^{t*}) > \epsilon$  or repeat  $I$  episodes;
    end
    Sample  $s'_1$  by acting updated  $\pi^l$  policy  $J$  steps;
     $s_1 \leftarrow s'_1$ 
end

```

that solves the navigation task as a high dimensional perception module (trained with automatically annotated data) and a low dimensional control module (trained with OIL).

A. Modular Architecture

The fundamental modules of our proposed system are summarized in Figure 2. We use a modular architecture [13, 24, 25] to reduce the required training time by introducing an intermediate representation layer. The overall neural network consists of two modules: a *Perception Network* ϕ and a *Control Network* φ . The input state includes image \mathcal{I} and the physical measurements \mathcal{M} of the vehicle's state (i.e. current orientation and velocity). The action is a control signal for the car (G : Gas/Brake, S : Steering) or UAV (T : Throttle, A : Aileron/Roll, E : Elevator/Pitch, R : Rudder/Yaw). One scalar value is regressed for each control signal. The perception network is parameterized by θ_p and the control network is parameterized by θ_c . The control network takes the intermediate representation predictions of the perception network and vehicle's state as input and outputs the final control predictions. In our experiments, we choose waypoints as the intermediate representation. Note that it can also be segmentation maps [24], depth images, affordances [5] [37] or the combination of them. The overall navigation policy can be described as follows:

$$\pi^l(s|\theta) = \pi^l(s|\theta_p, \theta_c) = \varphi(\phi(s_{\mathcal{I}}|\theta_p), s_{\mathcal{M}}|\theta_c) \quad (8)$$

The overall loss is defined in Equation 9 as a weighted sum of perception and control loss. The perception loss

relates to the intermediate layer by minimizing the difference between the ground truth intermediate representation z^* and predicted intermediate representation $\phi(s_{\mathcal{I}}|\theta_p)$. The control loss in Equation 10 comes from applying OIL to learn from multiple imperfect teachers (automated PID controllers in our case) by minimizing the surrogate loss in Equation 7, where $s \sim d_{\pi^t}$ and $a^* \sim \pi^{t*}$.

$$\mathbb{L} = \underbrace{\mathbb{L}_c(\pi(s|\theta), a^*)}_{\text{control loss}} + \lambda \underbrace{\mathbb{L}_p(\phi(s_{\mathcal{I}}|\theta_p), z^*)}_{\text{perception loss}} \quad (9)$$

$$\mathbb{L}_c(\pi(s|\theta), a^*) = \mathbb{L}_c(\varphi(\phi(s_{\mathcal{I}}|\theta_p), s_{\mathcal{M}}|\theta_c), a^*) \quad (10)$$

In general, this optimization problem can be solved by minimizing the overall loss with respect to θ_p and θ_c at the same time. The gradients are as follows:

$$\frac{\partial \mathbb{L}}{\partial \theta_p} = \frac{\partial \mathbb{L}_c}{\partial \theta_p} + \lambda \frac{\partial \mathbb{L}_p}{\partial \theta_p} = \frac{\partial \mathbb{L}_c}{\partial \varphi} \frac{\partial \varphi}{\partial \phi} \frac{\partial \phi}{\partial \theta_p} + \lambda \frac{\partial \mathbb{L}_p}{\partial \phi} \frac{\partial \phi}{\partial \theta_p} \quad (11)$$

$$\frac{\partial \mathbb{L}}{\partial \theta_c} = \frac{\partial \mathbb{L}_c}{\partial \varphi} \frac{\partial \varphi}{\partial \theta_c} \quad (12)$$

A good perception network is essential to achieve good controls. To maintain modularity and reduce training time, we first optimize only for θ_p and ignore the control loss. As the perception network converges, we fix θ_p and optimize for θ_c .

This modular approach has several advantages over an end-to-end approach (see also [13, 24, 25]). Since only the control module is specific to the vehicle’s dynamics, the perception module can simply be swapped out, allowing the vehicle to navigate in completely different environments without any modification to the control module. Similarly, if the reward function is changed in different tasks, we can simply retrain the control module to learn a different policy. Naturally, one can add links between the perception and control networks and then finetune the joint network in an end-to-end fashion. It is also possible to connect the two networks and use the waypoint labels as intermediate supervision for the perception part while training the joint model end-to-end. While these variants are interesting, we specifically refrain from such connections to safeguard the attractive modular properties.

In what follows, we provide details of the architecture, implementation, and training procedure for both the perception and control modules. Note that OIL and the proposed architecture can also be applied to other types of MDP problems (e.g. other vision-based sequential decision making problems).

B. Perception

In our case the perception module takes raw RGB images as an input and predicts a trajectory of waypoints relative to the vehicle’s current position, which remains unknown in 3D. The waypoints predicted by the perception module are input into the control network along with the current vehicle state (velocity and orientation).

Drawbacks of Predicting Controls Directly. Note that related work commonly proposed to frame the navigation task as an end-to-end learning problem, predicting controls directly from single images. However, this has several limitations:

(i) A teacher is required for data collection and the controls are strongly dependent on the teacher and vehicle dynamics, as opposed to our modular DNN. **(ii)** There is no unique mapping from images to controls, since different sequences of controls can lead to the same image. This can result in direct discrepancies and less stable training, if data from multiple teachers is used. **(iii)** When using camera views in addition to the views used for data acquisition (a common augmentation method to increase robustness), it is unclear how the controls corresponding to these augmented views should be adjusted, given the underlying complex nature of the task. For the case of driving, it might be sufficient to only predict a steering angle and acceleration. Since the car is confined to a 2D plane (road) and the friction of the tires limits drift, one can design a highly simplified model, e.g. by offsetting the steering by the rotation of the augmented camera view. However, for more complex scenarios where the vehicle moves in 3D and is less constrained by friction, such simplifying assumptions quickly become invalid.

Waypoint Encoding. The mapping from image to waypoints is deterministic and unique. For every camera view, the corresponding waypoints can easily be determined and are independent of the vehicle state. We define waypoints along the track as a vertical offset that is measured as the distance between the vehicle position and the projected point along the viewing axis, and a horizontal offset that is defined as the distance between the original and projected point along the viewing axis normal. We then encode these waypoints relative to the vehicle position and orientation by projecting them onto the viewing axis. Predicting waypoints rather than controls does not only facilitate network training, but it also allows for the automatic collection of training data without human intervention. Within the simulator, we simply sample/render the entire training track from multiple views and calculate the corresponding waypoints along the track. Note that it is still possible to use recordings from teachers, as one can use future positions to determine the waypoint for the current frame similar to [46]. Please refer to Figure 3 for a visualization of the encoding method.

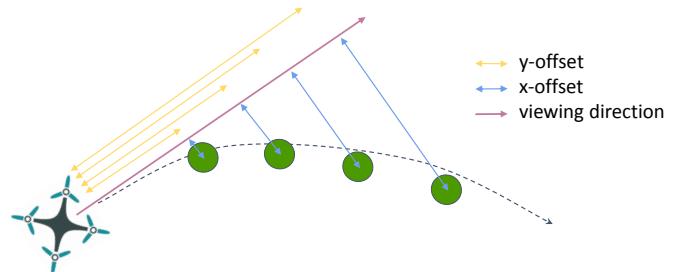


Fig. 3: Illustration of waypoint encoding.

Network Architecture. We choose a regression network architecture similar to the one used by Bojarski et al. [3]. Our DNN architecture is shown in Figure 2 as the perception module. It consists of eight layers: five convolutional

with $\{20, 24, 28, 30, 32\}$ filters and three fully-connected with $\{1800, 800, 100\}$ hidden units. The DNN takes in a single RGB-image with 180×320 pixel resolution and is trained to regress the next five waypoints (x -offset and y -offset with respect to the local position of the vehicle) using a standard ℓ_2 -loss and dropout ratio of 0.5 in the fully-connected layers. As compared to related methods [3, 40], we find that the relatively high input resolution is useful to improve the network's ability to look further ahead, and increasing long-term trajectory stability. For more details and a visualization of the learned perception network, please refer to the [appendix A](#).

C. OIL for Control

Here, we present details of the network architecture and learning strategy to train the control network using OIL.

Teachers and Learner. In our experiments, we use multiple naive PID controllers as teachers for our control policy. The PID parameters are tuned within a few minutes and validated on a training track to perform well. As the system is very robust to learn from imperfect teachers, we do not need to spend much effort tuning these parameters or achieve optimal performance of the teachers on all training tracks. Although an unlimited number of PID based teachers can be created, we empirically find five to be sufficient for the two control scenarios (autonomous driving and UAV racing). We refer to our evaluation section to see the effect of the number of teachers on learning. The five PID teachers have different properties. Some of them are fast at straightaways but are prone to crash in tight turns. Others drive or fly more conservatively and are precise at curves but slow at straightaways as a result. OIL enables the agent to learn from such imperfect teachers by selecting only the best maneuvers among them to learn.

We use a three-layer fully connected network to approximate the control policy φ of the learner. The MDP state of the learner is a vector concatenation of the predicted intermediate representation (e.g. waypoints) $\hat{z} = \phi(s_{\mathcal{I}}|\theta_p)$ and the vehicle state (physical measurements of the vehicle's speed and orientation) $s_{\mathcal{M}}$.

Network Architecture. The goal of the control network is to find a control policy φ that minimizes the control loss \mathbb{L}_c :

$$\varphi^* = \arg \min_{\varphi} \mathbb{L}_c(\varphi(\hat{z}, s_{\mathcal{M}}|\theta_c), a^*) \quad (13)$$

It consists of three fully-connected layers with hidden units $\{64, 32, 16\}$, where we apply dropout in the second layer with a ratio of 0.5 for regularization. The loss function \mathbb{L}_c is a standard ℓ_2 -loss optimized by the Adam optimizer with a learning rate of 10^{-4} . The control network is updated by OIL in an online fashion, while the vehicle runs through a set of training tracks.

As such, the control network is learning from experiences (akin to reinforcement learning), but it is supervised throughout by multiple teachers to minimize the surrogate loss. An advantage of our approach is that multiple teachers are able to teach throughout the control network's exploration. The control network never becomes dependent on the teachers,

but gradually becomes independent and eventually learns to outperform them.

V. EXPERIMENTS

A. Experimental Setup

The Sim4CV [23] (see Figure 1) environment provides capabilities to generate labeled datasets for offline training (e.g. imitation learning), and interact with the simulation environment for online learning (e.g. OIL or reinforcement learning) as well as online evaluation. For each application, we design six environments for training and four for testing. For fair comparison, human drivers/pilots are given as much time as needed to practice on the training tracks, before attempting the test tracks. Please refer to Figure 4 and Figure 5 for the test environments and to the [appendix A](#) for the training environments.

For the autonomous driving scenario, the PID teachers, learned baselines and human drivers have to complete one lap. They are scored based on the average error to the center line and the time needed to complete the course. In addition, they need to pass through invisible checkpoints placed every 50 meters to make sure they stay on the course and do not take any shortcuts. The vehicle is reset at the next checkpoint if it did not reach it within 15 seconds.

For the UAV racing task, all pilots have to complete two laps on the test tracks and are scored based on the percentage of gates they maneuver through and the overall time. Similar to the autonomous driving scenario we reset the UAV at the next checkpoint if it was not reached within 10 seconds. Here, the checkpoints correspond to the racing gates. Visualizations of the vehicle trajectory for all models on each track are provided in the [appendix A](#).

Training Details. For fair comparison we train each network until convergence or for at most 800k steps. We estimate the value functions for all the teachers and the learner to select the most critical teacher with N step rollouts ($N = 300$ for the car, $N = 200$ for the UAV) during the *observing* phase. During the *rehearsing* phase, we execute the learner policy for N steps until the advantage function $A(s_1, \pi^l, \pi^{t_*}) > \epsilon$ or $I = 50$ episodes are reached. We choose $\epsilon = -0.1V^{\pi^{t_*}}$ for the multi teacher experiments, and $\epsilon = 0$ for the single teacher experiments. We choose $J = 60$ as the *acting* step size.

Reward Function of OIL. The reward function of OIL is used to compute the value functions and advantage functions of learner and teachers. For OIL, we use reward functions that score the whole trajectory. Assuming a trajectory of length N , let $\tau = (s_1, a_1, s_2, a_2, \dots, s_N)$ denote the generated trajectory and S_{end} denote all the terminal states when the vehicles leave the course/track or hit an obstacle.

For autonomous driving, preciseness in lane keeping is an important measurement that reflects the smoothness and safety of the algorithm. Hence, we use a reward function that considers both speed and trajectory error as follows:

$$R(\tau) = \begin{cases} \frac{\zeta}{\alpha \sum_{n=1}^N e_n + 1} & s_N \notin S_{\text{end}} \\ \frac{\zeta}{\alpha \sum_{n=1}^N e_n + 1} + r_{\text{penalty}} & s_N \in S_{\text{end}} \end{cases} \quad (14)$$

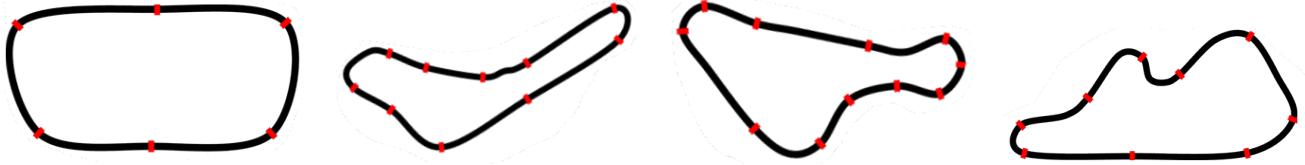


Fig. 4: Tracks used to evaluate the flying policy.



Fig. 5: Maps used to evaluate the driving policy.

where ζ is the trajectory length of the vehicle along the center of the lane and e_n is the distance error to the center of the lane at state s_n . We choose $\alpha = 0.5$ and a large negative reward $r_{\text{penalty}} = -15000$ for violations (e.g. crossing the lane boundaries).

For UAV racing, it is not necessary to stay in the center of the track. Hence, we use the following reward function that considers forward speed v_f and terminal state S_{end} :

$$R(\tau) = \begin{cases} \sum_{n=1}^N v_{f_n} & s_N \notin S_{\text{end}} \\ \sum_{n=1}^N v_{f_n} + r_{\text{penalty}} & s_N \in S_{\text{end}} \end{cases} \quad (15)$$

where v_{f_n} is the forward speed of the vehicle after executing a_n at state s_n and r_{penalty} is a large negative reward. We choose the penalty to be -15000 .

Reward function of DDPG. In our experiments, DDPG cannot converge using the same simple reward function as OIL. Therefore, we design a more complex reward function to induce a dense normalized reward as follows.

$$r_{s_n, a_n} = \begin{cases} \frac{e_n - e_{n+1}}{e_{\text{norm}}} + \frac{1}{e_{n+1} + 1} + \beta \frac{v_{f_n}}{v_{\text{norm}}} & s_{n+1} \notin S_{\text{end}} \\ r_{\text{penalty}} & s_{n+1} \in S_{\text{end}} \end{cases} \quad (16)$$

where e_n and e_{n+1} are the distance errors to the center of the track/lane at state s_n and state s_{n+1} respectively, e_{norm} is a distance error normalization factor, v_{f_n} is the forward speed the vehicle, v_{norm} is a velocity normalization factor and r_{penalty} is a negative reward for when the vehicles reaches a terminal state. Furthermore, we find it is hard to learn both acceleration and steering controls for the car. Therefore, we fix the acceleration value to $E = 0.5$ and choose $\beta = 0$ to only score the agent by the distance error. For the UAV, we choose $\beta = 1$ to score the agent by both the velocity and distance error and choose $v_{\text{norm}} = 800$. We set $e_{\text{norm}} = 15$ and $r_{\text{penalty}} = -0.2$ for both the UAV and the car.

Controller Architecture of Learned Baselines. For fair comparisons, we use the same control policy network with hidden units $\{64, 32, 16\}$ for all the learned baselines. As for the critic network of DDPG, we use the same number of hidden units and use an output layer with a one-dimensional output to predict the Q value.

B. Results

Comparison to State-of-the-Art Baselines. We compare OIL for both autonomous driving and UAV racing to several DNN baselines. These include both IL approaches (Behaviour Cloning, DAGGER) and RL approaches (DDPG). For each comparison we implement both learning from the single best teacher and an ensemble of teachers. This essentially allows a broad baseline comparison of 6 different state-of-the-art learning approaches evaluating various IL and RL approaches.

In Table I, the learned approaches are compared to OIL. Our results demonstrate that OIL outperforms all learned baselines in both accuracy scores and timings. Behaviour Cloning and DAGGER both perform worse in terms of score and timings as the number of teachers increases. This is because they do not distinguish between good and bad behaviour and learn from both. Moreover, they are limited by their teachers and unable to outperform them. This can be seen in single teacher training, where they do not achieve scores better than their teacher. In contrast, OIL improves upon scores both with a single teacher and multiple teachers. In comparison to DDPG, OIL converges quickly without extensive hyperparameter search and still learns to fly/drive much more precisely and faster.

Comparison to Teachers and Human Performance. We compare our OIL trained control network to the teachers it learned from and human performance. The perception network is kept the same for all learned models. The summary of this comparison to OIL is given in Table I. Our results demonstrate

Results for the UAV			Results for the car		
Teachers: PID controllers			Teachers: PID controllers		
Teacher #1	100%	131.9	Teacher #1	24.1	151.4
Teacher #2	76.4%	87.0	Teacher #2	539.6	110.0
Teacher #3	97.2%	87.6	Teacher #3	19.5	84.5
Teacher #4	80.6%	90.3	Teacher #4	76.7	76.7
Teacher #5	69.4%	99.7	Teacher #5	568.4	102.2
Baseline: Human			Baseline: Human		
Novice	97.2%	124.6	Novice	85.3	100.7
Intermediate	100%	81.2	Intermediate	80.6	88.3
Expert	100%	46.9	Expert	49.0	70.4
Learned Policy: Best Teacher (1)			Learned Policy: Best Teacher (3)		
Behaviour Cl.	94.4%	139.6	Behaviour Cl.	13.9	88.6
DAGGER	100%	134.1	DAGGER	38.6	88.3
DDPG	95.8%	84.6	DDPG	57.4	139.6
OIL (single)	100%	133.9	OIL (single)	12.4	88.5
Learned Policy: All Teachers			Learned Policy: All Teachers		
Behaviour Cl.	72.2%	101.6	Behaviour Cl.	388.6	112.7
DAGGER	58.3%	140.1	DAGGER	25.9	87.8
DDPG	95.8%	84.6	DDPG	57.4	139.6
OIL (multi)	100%	81.3	OIL (multi)	17.6	74.2

TABLE I: *Left:* Results for the UAV. Columns show the number of gates passed and time to complete two laps. *Right:* Results for the car. Columns show average error to center of the road and time to complete one round. All results are averaged over all 4 test tracks/maps. Please refer to the [appendix A](#) for detailed results per track/map.

Ablation study for the car		
OIL (Teachers #1-5, 60 steps)	15.3	80.5
OIL (Teachers #1-5, 180 steps)	16.0	80.7
OIL (Teachers #1-5, 300 steps)	17.6	74.2
OIL (Teachers #1-5, 600 steps)	13.8	82.2
OIL (Teachers #1,3,4, 300 steps)	25.8	73.6

TABLE II: Ablation study for the car. *Left column:* average error to center of the road. *Right column:* time to complete one round. All results are averaged over all 4 test tracks/maps. Please refer to the [appendix A](#) for the detailed results per track/map.

that OIL outperforms all teachers and novice to intermediate human pilots/drivers.

Teacher #1 has the least error of all teachers and is the only one to perfectly complete all gates in the UAV racing evaluation. However, OIL not only completes all gates but is also much faster when learning from multiple teachers. In the autonomous driving evaluation, it is more precise in centering along the middle of the road and also much faster when learning from multiple teachers. In comparison to humans, OIL is better than novice and intermediate levels but slower than an expert in both applications. A notable difference between the expert driver and OIL is that OIL has a much lower error in driving. It is able to maintain high speeds while staying most accurately in the center of the tracks. Compared to the expert driver, OIL is 3.83 seconds slower but only has 35.91% of the error in terms of the distance to the center.

Ablation Study. We investigate the importance of trajectory length and the number of teachers, and report results in Table II. Our experiments show that OIL is robust to different trajectory lengths and varying numbers of teachers. One observation is that different trajectory length balances the average error and speed. In our car experiment, when we set trajectory length to 300 steps and use five teachers, OIL learns a fast and precise (compared to human experts and teachers) driving policy. Moreover, it is observed that increasing the number of teachers reduces the average error and makes the learned policy more stable. When OIL only learns from three teachers, the time to complete one round reduces slightly but the average error increases significantly compared to learning from five teachers.

VI. CONCLUSION

In this paper, we present Observational Imitation Learning (OIL), a new approach for training a stationary deterministic policy that is not bound by imperfect or inexperienced teachers since its policy is updated by selecting only the best maneuvers at different states. OIL can be regarded as a generalization of DAGGER for multiple imperfect teachers.

We demonstrate the ability of OIL by training a control network to autonomously drive a car and to fly an unmanned aerial vehicle (UAV) through challenging race tracks. Extensive experiments demonstrate that OIL outperforms single and multiple-teacher learned IL methods (Behavior Cloning, DAGGER), RL approaches (DDPG), its teachers and experienced humans pilots/drivers.

While our method works well in the navigation tasks with access to non-expert controllers, a question arises concerning how this method would scale to more complex tasks where controller design is very time-consuming or intractable, e.g. Atari games. OIL is general in nature and only requires access to teachers and a function to score them; one approach to avoid hand-designed controllers would be to use learned ones instead (e.g. trained using different RL algorithms) and use the reward of the environment as a score for OIL.

VII. FUTURE WORK

OIL provides a new learning-based approach that can replace traditional control methods especially in robotics and control systems. We expect OIL to expand its reach to other areas of autonomous navigation (e.g. obstacle avoidance) and benefit other robotic tasks (e.g. visual grasping or placing).

One very interesting avenue for future work is to apply OIL outside simulation. Although Sim4CV uses a high fidelity game engine for rendering and physics, the differences between the simulated world and the real world will need to be reconciled. Real-world physics, weather and road conditions, as well as sensor noise will present new challenges for adapting/training the control network.

ACKNOWLEDGMENTS

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding.

REFERENCES

- [1] Olov Andersson, Mariusz Wzorek, and Patrick Doherty. Deep learning quadcopter control via risk-aware active learning. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI), 2017, San Francisco*, 2017. [2](#)
- [2] Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A. Knepper, and Yoav Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania*, 2018. [2](#)
- [3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016. [2, 5, 6](#)
- [4] Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *ICML*, 2015. [2](#)
- [5] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015. [2, 4](#)
- [6] Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings*, 2017. [1, 2](#)
- [7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 13–15 Nov 2017. [2](#)
- [8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR 2016*, pages 4340–4349, 2016. [2](#)
- [9] Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, Workshop Track Proceedings*, 2018. [2](#)
- [10] Michael Gimelfarb, Scott Sanner, and Chi-Guhn Lee. Reinforcement learning with multiple experts: A bayesian model combination approach. In *Advances in Neural Information Processing Systems 31*, pages 9549–9559. 2018. [2](#)
- [11] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pages 4565–4573. 2016. [2](#)
- [12] David Isele, Reza Rahimi, Akansel Cosgun, Kaushik Subramanian, and Kikuo Fujimura. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 2034–2039, 2018. [2](#)
- [13] Elia Kaufmann, Antonio Loquercio, Rene Ranftl, Alexey Dosovitskiy, Vladlen Koltun, and Davide Scaramuzza. Deep drone racing: Learning agile flight in dynamic environments. In *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 133–145, 2018. [2, 4, 5](#)
- [14] Jan Koutník, Jürgen Schmidhuber, and Faustino Gomez. *Online Evolution of Deep Convolutional Network for Vision-Based Reinforcement Learning*, pages 260–269. Springer International Publishing, Cham, 2014. doi: 10.1007/978-3-319-08864-8_25. [2](#)
- [15] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *ICML 2016, New York City, NY*, pages 430–438, 2016. [2](#)
- [16] Sergey Levine and Vladlen Koltun. Guided policy search. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, Atlanta, Georgia, 2013. [2](#)
- [17] Yuxiang Li, Ian Kash, and Katja Hofmann. Learning good policies from suboptimal demonstrations. In *14th European Workshop on Reinforcement Learning (EWRL 2018)*, October 2018. [2](#)
- [18] Xiaodan Liang, Tairui Wang, Luona Yang, and Eric Xing. CIRL: controllable imitative reinforcement learning for vision-based self-driving. In *ECCV 2018, 15th European Conference, Munich, Germany*, pages 604–620, 2018. [2](#)
- [19] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *ICLR*, abs/1509.02971, 2016. [2, 3](#)
- [20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015. [3](#)
- [21] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016. [2, 3](#)
- [22] M. Müller, V. Casser, N. Smith, D. L. Michels, and B. Ghanem. Teaching UAVs to Race: End-to-End Regression of Agile Controls in Simulation. In *European Conference on Computer Vision Workshop (ECCVW)*, September 2018. [2](#)
- [23] Matthias Müller, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem. Sim4cv: A photo-realistic simulator for computer vision applications. *Int. J. Comput. Vision*, 126(9):902–919, September 2018. ISSN 0920-5691. doi: 10.1007/s11263-018-1073-7. [1, 6](#)
- [24] Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. Driving policy transfer via modularity and abstraction. In *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of*

- Machine Learning Research*, pages 1–15. PMLR, 29–31 Oct 2018. 4, 5
- [25] Matthias Müller, Guohao Li, Vincent Casser, Neil Smith, Dominik L. Michels, and Bernard Ghanem. Learning a controller fusion network by online trajectory filtering for vision-based uav racing. In *CVPR Workshops*, June 2019. 2, 4, 5
- [26] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999. 1
- [27] Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile autonomous driving using end-to-end deep imitation learning. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. 2
- [28] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In *European Conference on Machine Learning*, pages 280–291. Springer, 2005. 3
- [29] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Robotics: Science and Systems (RSS)*, 2018. 2
- [30] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2
- [31] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2
- [32] Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668, 2010. 3
- [33] Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014. 2
- [34] Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. No-regret reductions for imitation learning and structured prediction. *CoRR*, abs/1011.0686, 2010. 2, 3
- [35] Ahmad El Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. End-to-end deep reinforcement learning for lane keeping assist. *CoRR*, abs/1612.04340, 2016. 2
- [36] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017. ISSN 2470-1173. doi: doi:10.2352/ISSN.2470-1173.2017.19.AVM-023. 2
- [37] Axel Sauer, Nikolay Savinov, and Andreas Geiger. Conditional affordance learning for driving in urban environments. In *Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland*, pages 237–252, 2018. 4
- [38] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897, 2015. 2
- [39] Shital Shah, Debadeepa Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics, Results of the 11th International Conference, FSR 2017, Zurich, Switzerland, 12-15 September 2017*, pages 621–635, 2017. 2
- [40] Nikolai Smolyanskiy, Alexey Kamenev, Jeffrey Smith, and Stan Birchfield. Toward low-flying autonomous MAV trail navigation using deep neural networks for environmental awareness. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 4241–4247, 2017. 2, 6
- [41] Riccardo Spica, Davide Falanga, Eric Cristofalo, Eduardo Montijano, Davide Scaramuzza, and Mac Schwager. A real-time game theoretic planner for autonomous two-player drone racing. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi: 10.15607/RSS.2018.XIV.040. 2
- [42] Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply aggravated: Differentiable imitation learning for sequential prediction. In *ICML 2017, Sydney, NSW, Australia*, 2017. 2
- [43] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998. 2, 3
- [44] Matthew E. Taylor, Halit Bener Suay, and Sonia Chernova. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2, AAMAS'11, Richland, SC, 2011*. 2
- [45] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4950–4957, 2018. 2
- [46] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *CoRR*, abs/1612.01079, 2016. 5
- [47] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3530–3538, 2017. 2
- [48] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450*, 2016. 2
- [49] Yuke Zhu, Ziyu Wang, Josh Merel, Andrei A. Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, Jnos Kramr, Raia Hadsell, Nando de Freitas, and Nicolas Heess. Reinforcement and imitation learning for diverse visuomotor skills. In *Robotics: Science and Systems*, 2018. 2

APPENDIX

Perception Network Visualization. After training the perception network to convergence, we visualize how parts of the network behave in order to get additional insights on what it has intrinsically learned. Figures 6 and 7 show some feature maps for each convolutional layer of the trained networks for an input image. Note how the learned filters are able to extract all necessary information in the scene (i.e. road boundaries or gates and cones). While filters from early layers already produce a reasonable activation heatmap, the deeper layers abstract the relevant features even more clearly. Although the feature map resolution becomes very low in the deeper layers, high activations can still be identified for parts of the image with semantic meaning for autonomous driving and UAV racing.

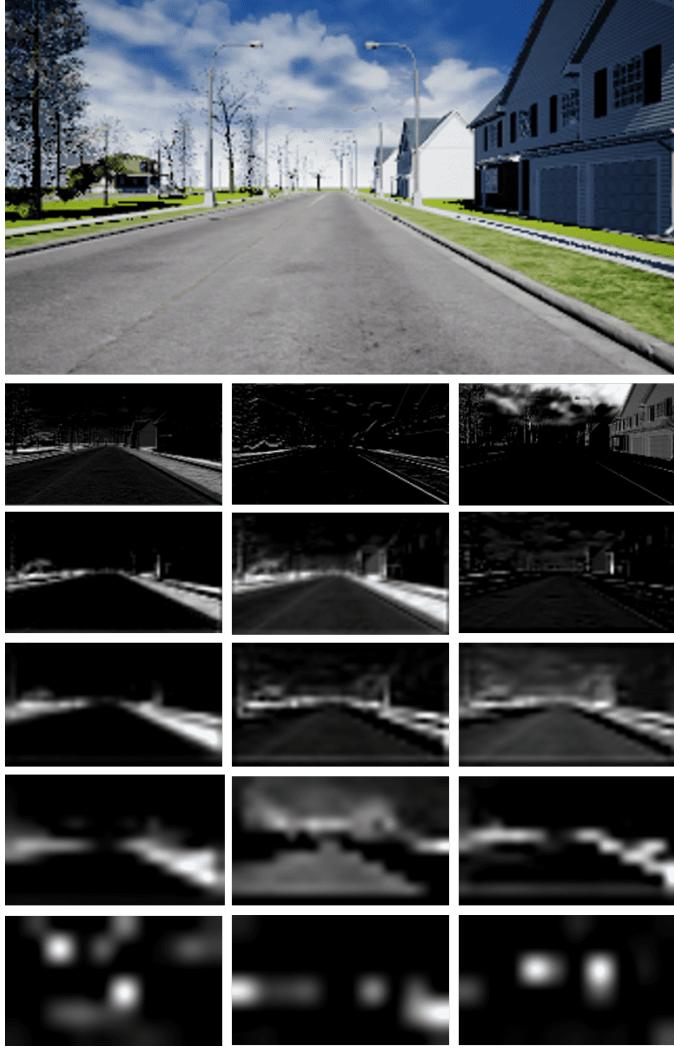


Fig. 6: Visualization of selected feature maps from convolutional layers 1 to 5 (top to bottom) in our trained perception network. Top image is input at 180×320 resolution. Notice how the network activates in locations of semantic meaning for the task of autonomous driving, namely the road boundaries.

Detailed Experimental Results. We show detailed results for the autonomous driving experiments for each course and report the average error to the center of the lane and the time to complete the course in Tables III, IV and V. In Table III, we show the results for each of the teachers and provide human baselines. In Table IV, we compare our approach *OIL* to several learned baselines. In Table V, we show an ablation study of *OIL* in regard to trajectory length and number of teachers. Figures 8, 9, 10 and 11 visualize the results.

We show detailed results for the autonomous UAV racing experiments for each track and report the number of missed gates and the time to complete two laps in Tables VI and VII. In Table VI, we show the results for each of the teachers and provide human baselines. In Table VII, we compare our approach *OIL* to several learned baselines. Figures 12, 13, 14 and 15 visualize the results.

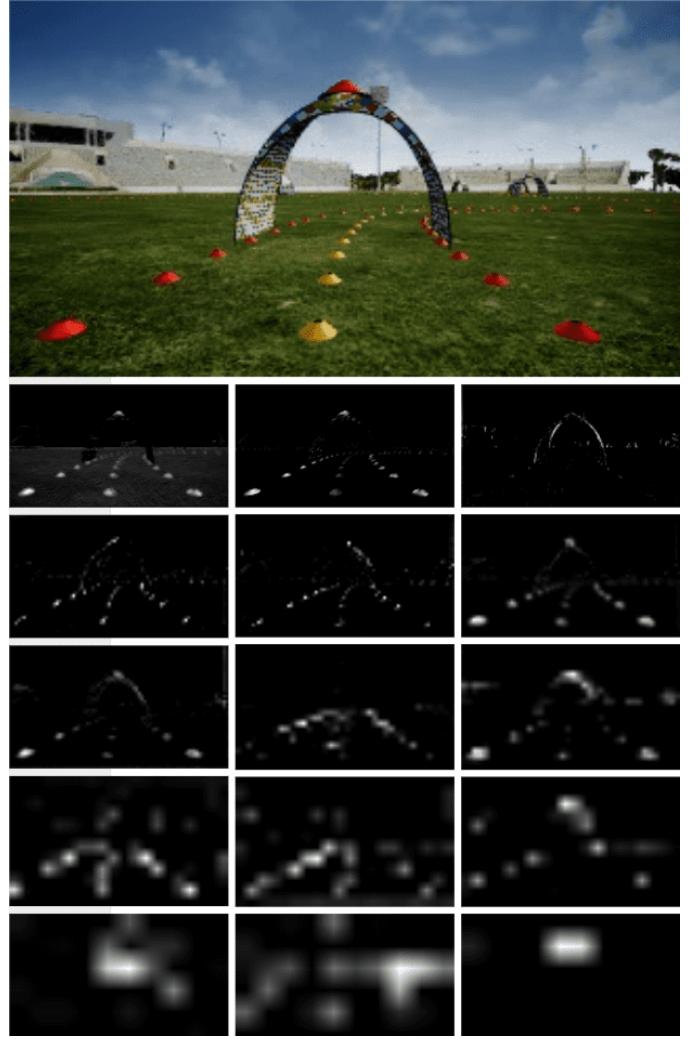


Fig. 7: Visualization of selected feature maps from convolutional layers 1 to 5 (top to bottom) in our trained perception network. Top image is input at 180×320 resolution. Notice how the network activates in locations of semantic meaning for the task of UAV racing, namely the gates and cones.

Transfer. The modular network architecture abstracts perception and control, and can be applied to many different types of complex vision-based sequential prediction problems. The perception network can be specific to the environment making a transfer of the learnt controller (e.g. sim-to-real) easier but predicted trajectories and state estimation will still need to be consistent across environments. The transferability of the perception network can be tested with a conservative PID controller handling controls. The transferability of the control network can be tested using an IMU and differential GPS to determine waypoints and state.

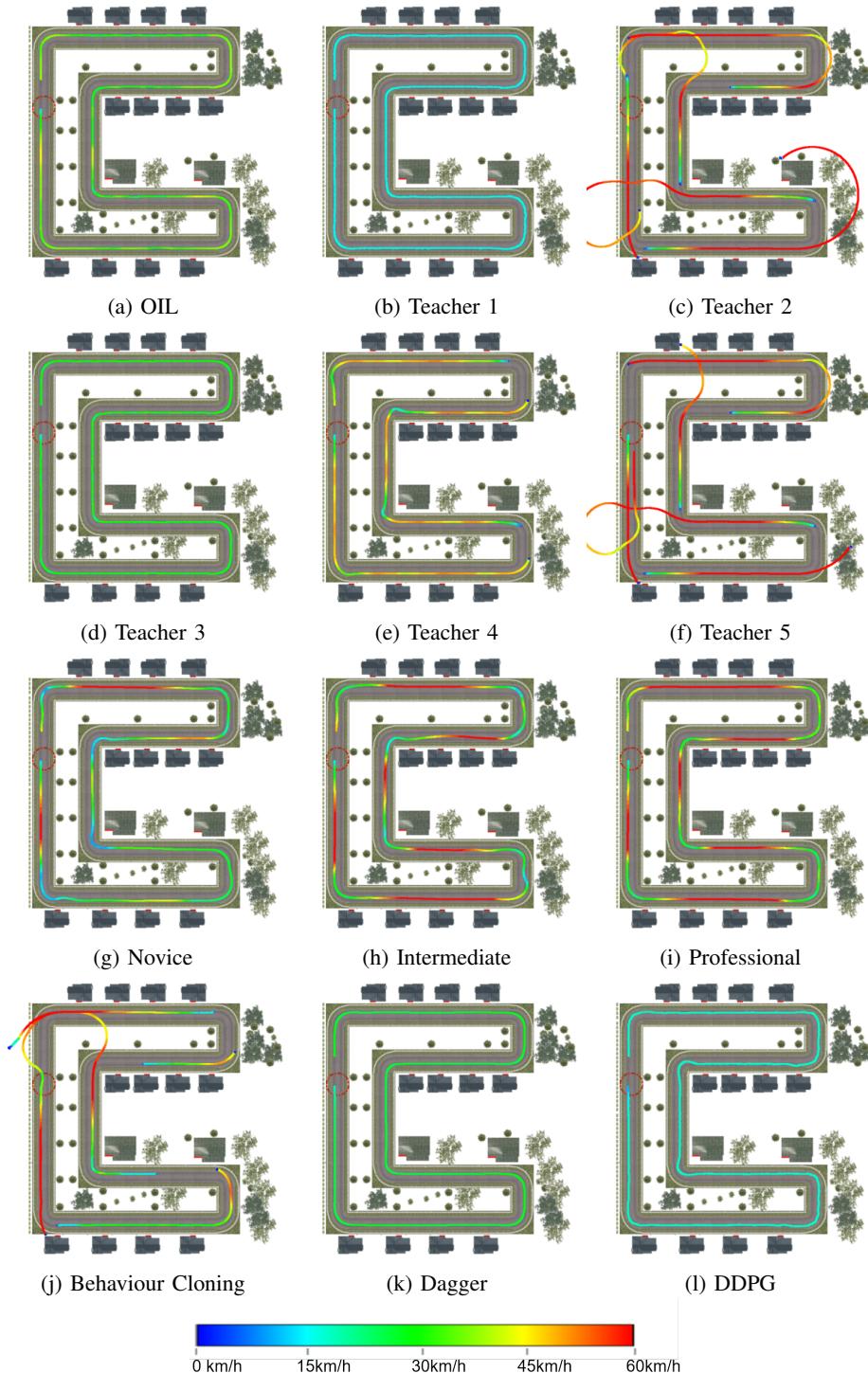


Fig. 8: Comparison between our learned policy and its teachers (row 1,2), human drivers (row 3) and baselines (row 4) on test course 1. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum speed.

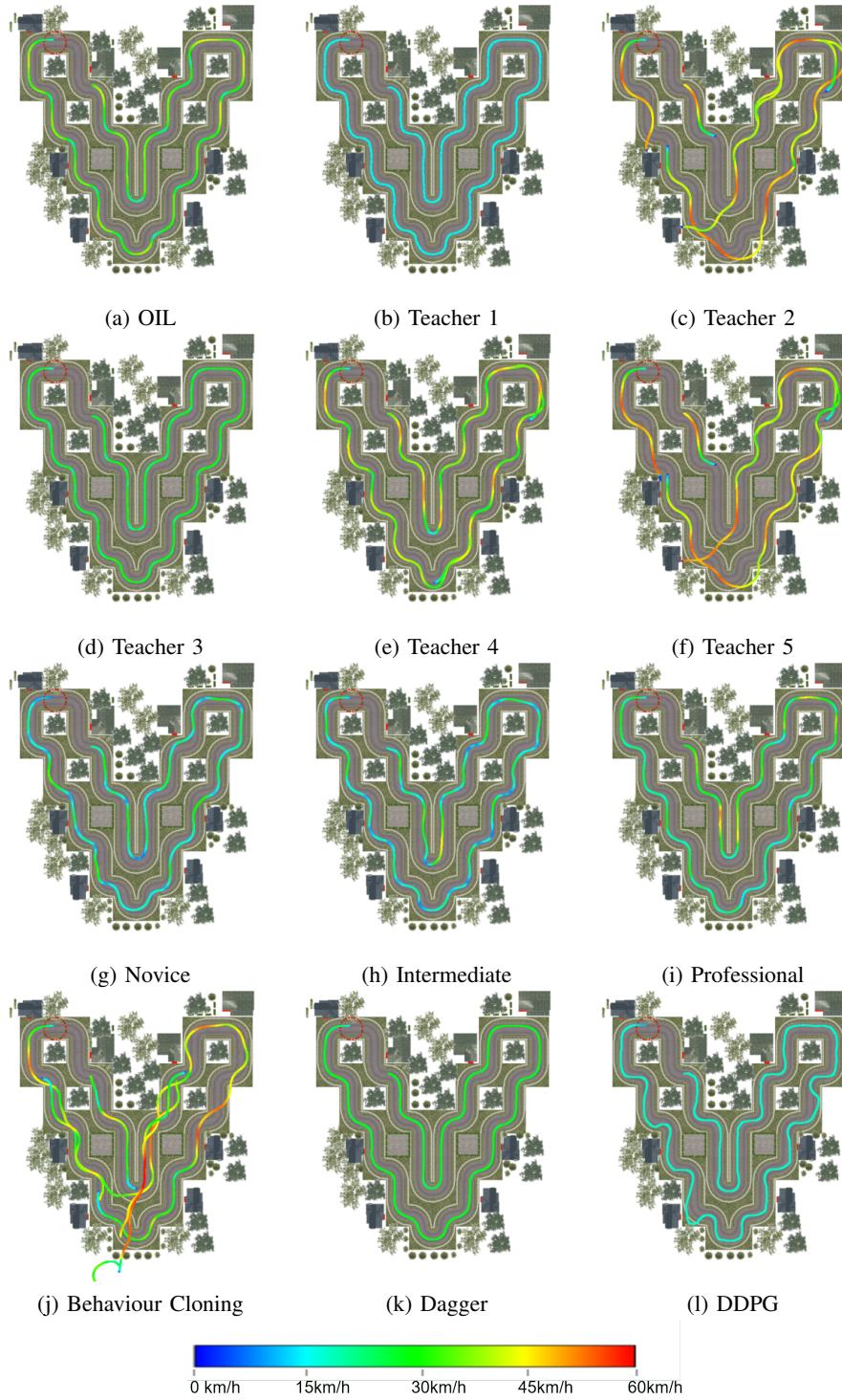


Fig. 9: Comparison between our learned policy and its teachers (row 1,2), human drivers (row 3) and baselines (row 4) on test course 2. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum speed.

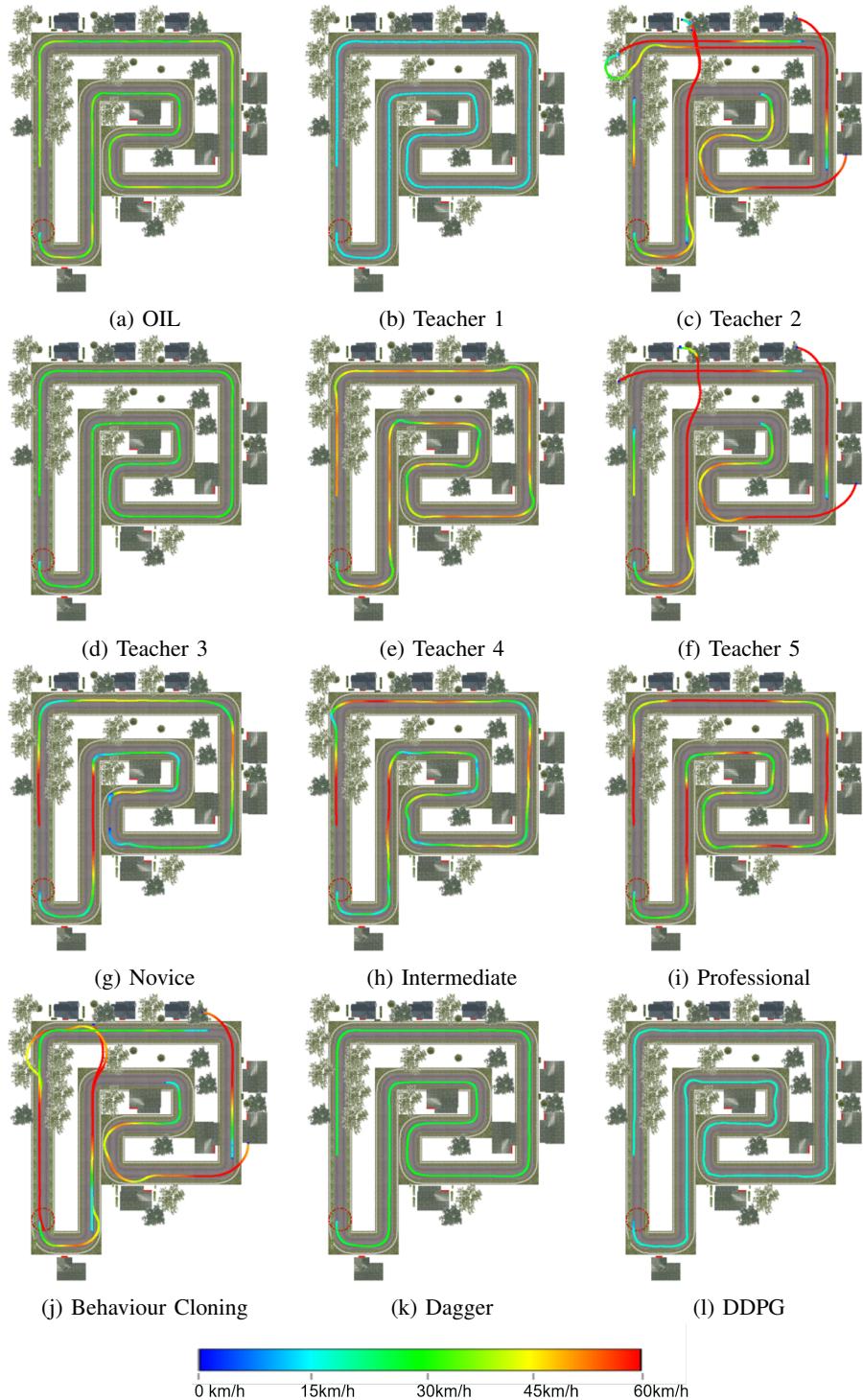


Fig. 10: Comparison between our learned policy and its teachers (row 1,2), human drivers (row 3) and baselines (row 4) on test course 3. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum speed.

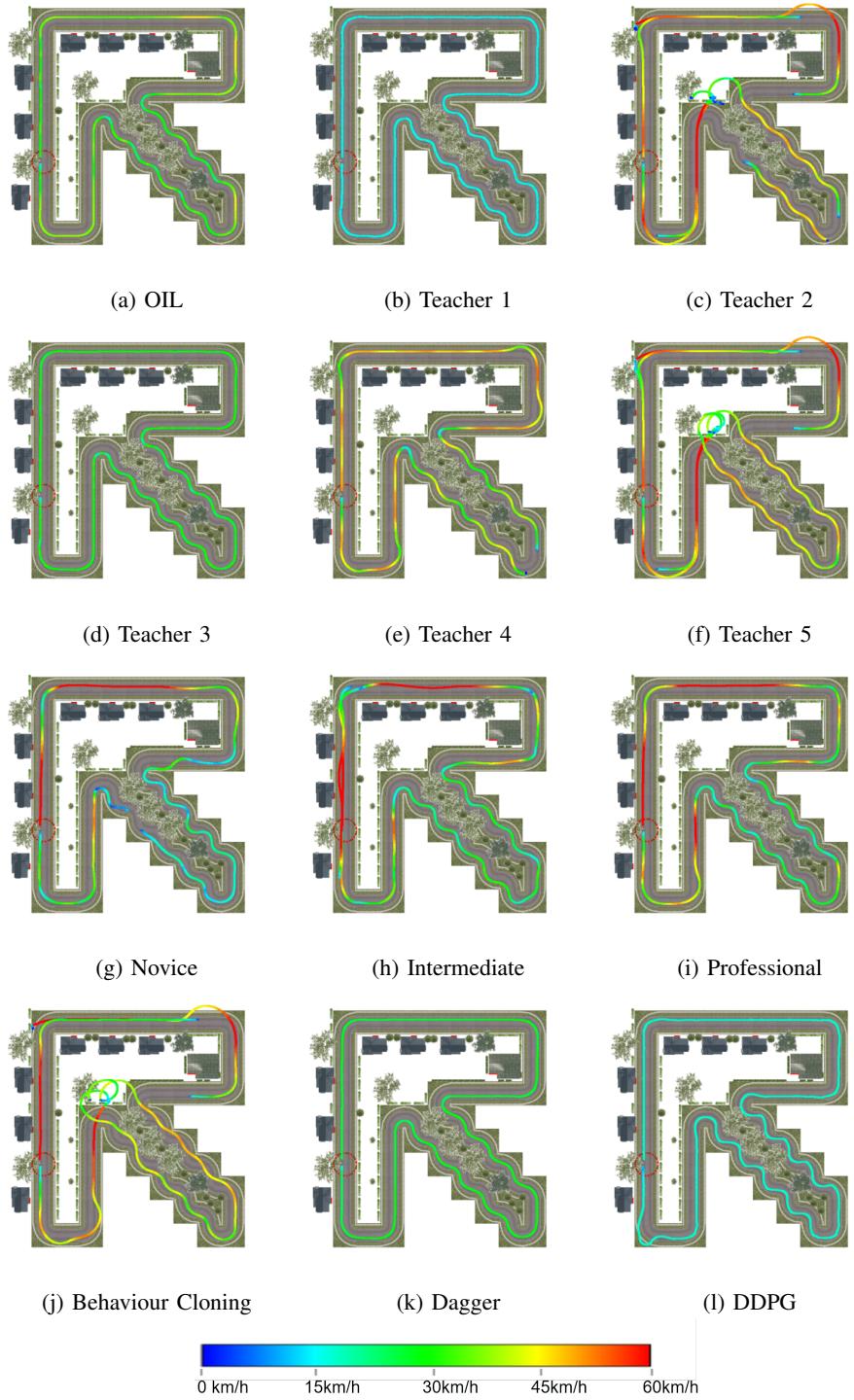


Fig. 11: Comparison between our learned policy and its teachers (row 1,2), human drivers (row 3) and baselines (row 4) on test course 4. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum speed.

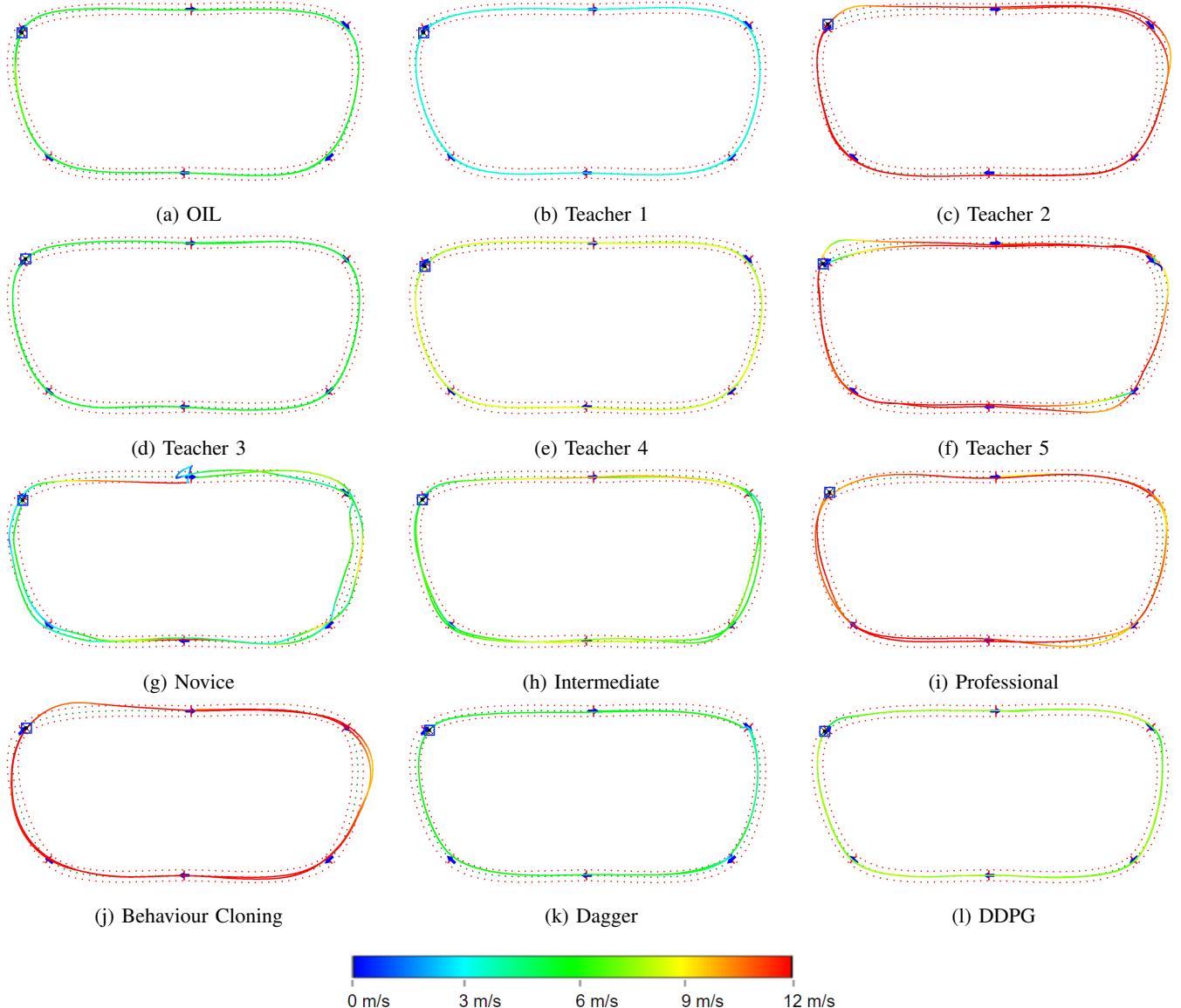


Fig. 12: Comparison between our learned policy and its teachers (row 1,2), human pilots (row 3) and baselines (row 4) on test track 1. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum.

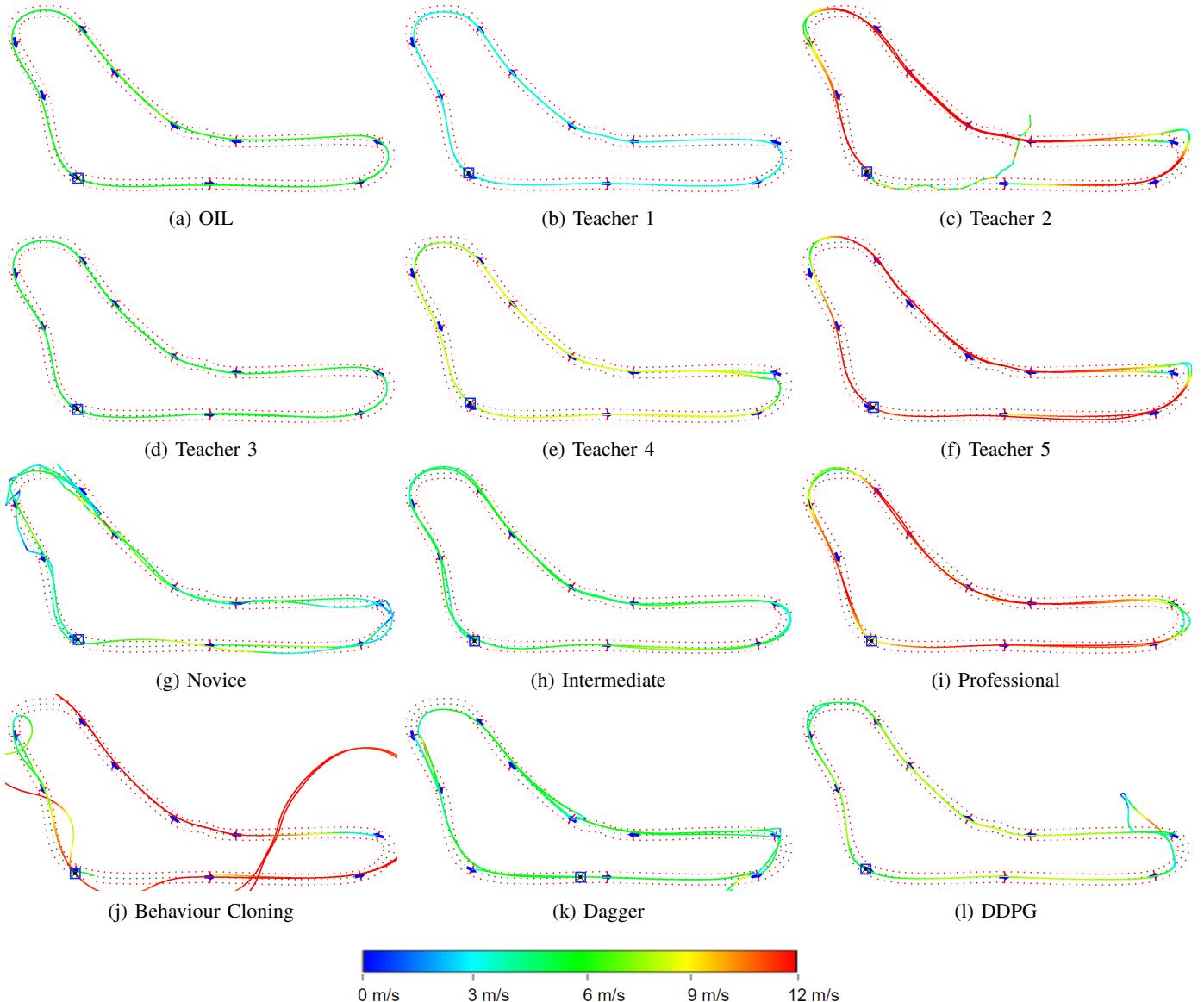


Fig. 13: Comparison between our learned policy and its teachers (row 1,2), human pilots (row 3) and baselines (row 4) on test track 2. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum.

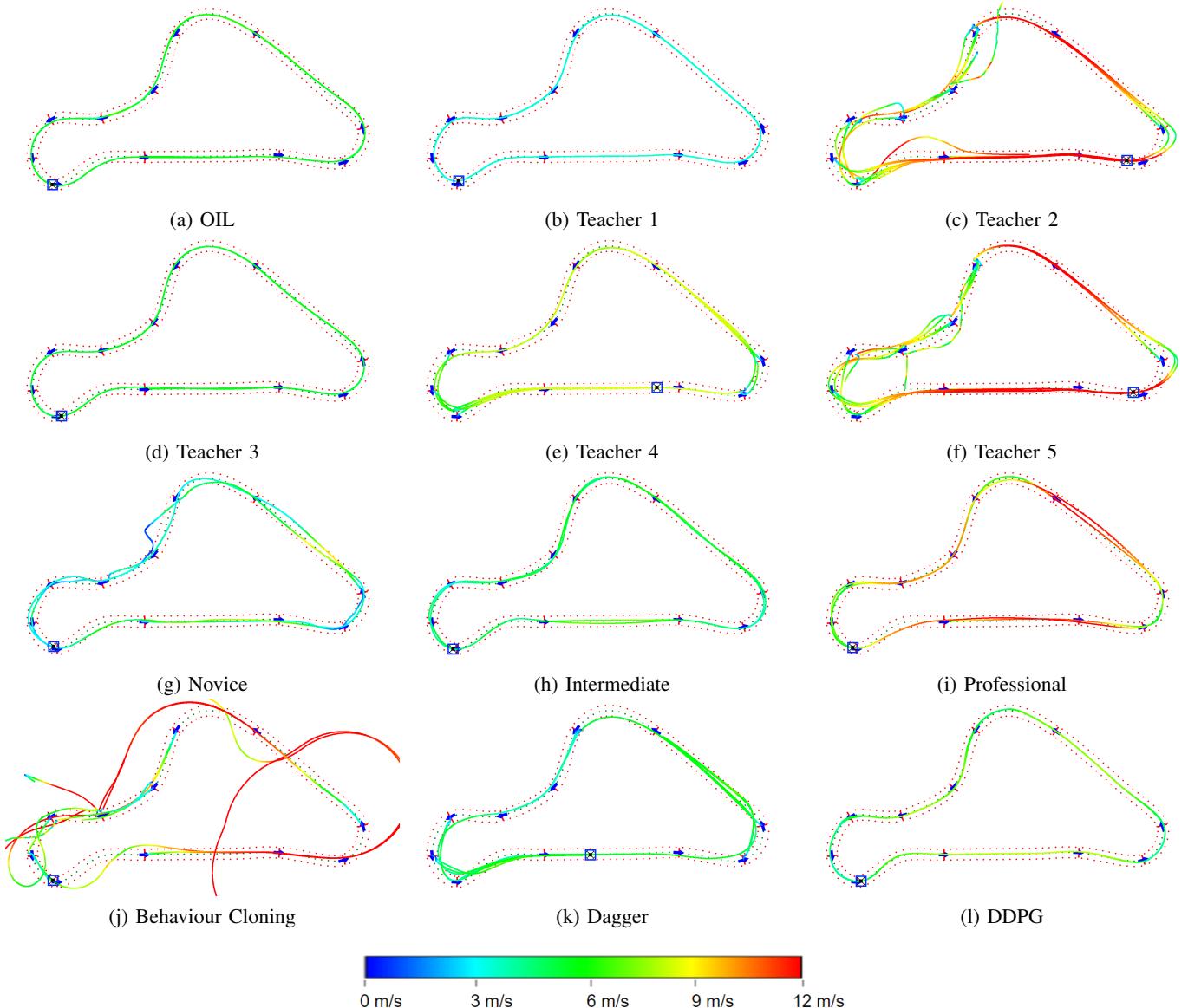


Fig. 14: Comparison between our learned policy and its teachers (row 1,2), human pilots (row 3) and baselines (row 4) on test track 3. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum speed.

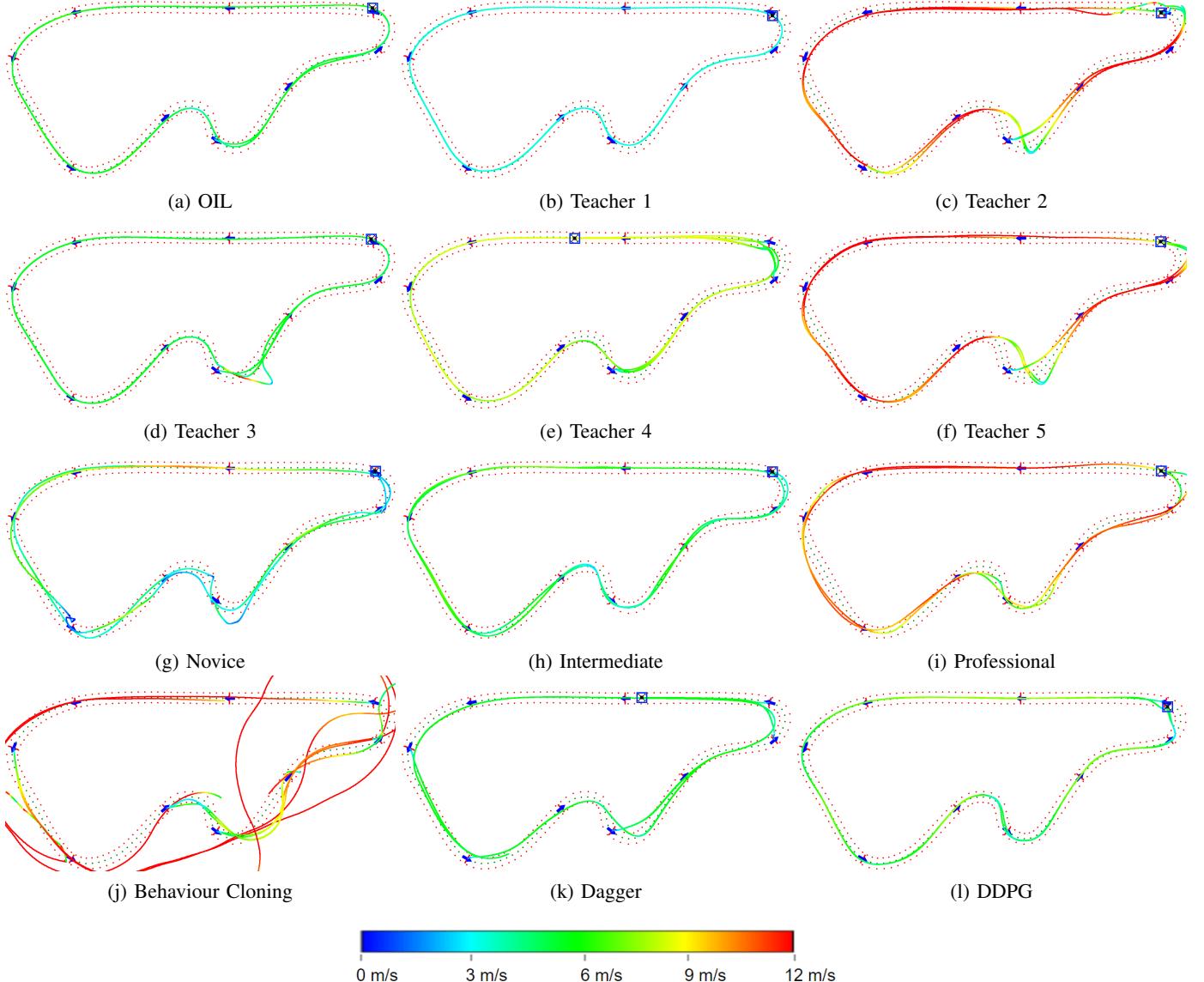


Fig. 15: Comparison between our learned policy and its teachers (row 1,2), human pilots (row 3) and baselines (row 4) on test track 4. Color encodes speed as a heatmap, where blue is the minimum speed and red is the maximum.