

# Vision-based Control of a Quadrotor in User Proximity: Mediated vs End-to-End Learning Approaches

Dario Mantegazza, Jérôme Guzzi, Luca M. Gambardella and Alessandro Giusti

**Abstract**—We consider the task of controlling a quadrotor to hover in front of a freely moving user, using input data from an onboard camera. On this specific task we compare two widespread learning paradigms: a mediated approach, which learns a high-level state from the input and then uses it for deriving control signals; and an end-to-end approach, which skips high-level state estimation altogether. We show that despite their fundamental difference, both approaches yield equivalent performance on this task. We finally qualitatively analyze the behavior of a quadrotor implementing such approaches.

VIDEOS, DATASETS, AND CODE

Videos, data, and code to reproduce our results are available at: <https://github.com/idsia-robotics/proximity-quadrotor-learning>.

## I. INTRODUCTION

Robot control systems are traditionally structured in two distinct modules: perception and control. Perception processes the robot’s input in order to derive a high-level state, which represents meaningful and relevant information for the task the robot needs to solve. State information is then used by a controller, which determines the low-level control signals to be provided to the hardware.

For mobile robots operating in real-world unstructured environments, perception is often challenging. This is especially true when this step involves the interpretation of complex, high-dimensional data such as images. Many recent successful systems deal with this problem by adopting supervised machine learning (ML) techniques which operate on sensing data as input.

When designing such a system, we face the choice between at least two approaches.

- A mediated approach, in which one trains a supervised ML model to predict the high-level state given the inputs; then, control signals are derived from the state using a designed controller (or another learned model).
- An end-to-end approach, in which a supervised ML model is trained to directly predict the control signals from the sensing data, without passing through an intermediate high-level representation of the state.

Which architecture is preferable? In some situations, one might not have a choice: for example, if one can not (or does not want to) collect ground truth information about the high-level state, training a perception model for a mediated approach is not possible (and this fact motivates many

end-to-end systems). However, sometimes datasets annotated with such a ground truth can be acquired, potentially at a cost. Is it worth it?

There are several advantages in a mediated approach: 1) in many mobile robotics applications, hand-designing a controller, given a high-level state representation, is feasible and gives the designer explicit control on the resulting robot behavior; 2) a mediated approach is more transparent and may be easier to inspect and debug: given an unexpected robot behavior, the designer can inspect the high-level state to determine whether the problem is in the perception or controller.

In contrast, end-to-end approaches are appealing because they are conceptually very simple and can potentially be more computationally efficient, especially if the high-level state representation is complex and high-dimensional. Moreover, as mentioned above, end-to-end approaches don’t depend on high-level state’s ground truth for training.

The considerations above disregard one key issue, i.e., the difficulty of learning models for the two approaches. Is learning a perception model (which outputs high-level state) easier than learning an end-to-end model (which directly outputs control signals)? Does one of the two models require a larger amount of training data to reach the same performance?

This is a key issue to consider when designing ML-based robot controllers: we investigate this question for one specific task. After reviewing related literature (Section I-A), we model and formalize the end-to-end architecture and two variants of the mediated architecture (Section II-A); we instantiate these three architectures for one specific task, that we consider in the remainder of the paper, i.e., controlling a quadrotor in order to hover in front of a person who is freely moving (Section II); such behavior could be implemented by a quadrotor tasked to monitor a person or expecting commands from them. The **main contribution** of the paper is a set of experiments showing that, for this specific task, training models for the three architectures has the same difficulty (the setup and results for these experiments are described in Sections III and IV respectively). The **main limitation** is that we limit our study to a single reactive control task; nonetheless, it is a challenging, real-world task which has several characteristics in common with other important subproblems in mobile robotics. A **secondary contribution** of this paper is the design, implementation and validation of the drone control system described above, which constitutes a useful component in applications involving proximal interaction of humans and quadrotors; collected datasets, source code and training models are available for

download.

### A. Related Work

End-to-end learning approaches map raw sensor data (lidar, camera images, ego-motion sensors) to control actions: either as direct low-level control outputs [1], or as target points (e.g., desired velocity) for optimal low-level controllers [2].

Supervised end-to-end approaches have been successfully applied to a variety of challenging control scenarios: off-road obstacle avoidance [1], [3]; autonomous driving [4], [5], [6], [7], [8]; vision-based manipulation [9]; quadrotor control in forested environments [10], [11], [12], cities [13], [14], and cluttered environments [2].

These can be considered instances of imitation learning (also known as learning-from-demonstration): in fact, one trains a policy to solve a sequential task using only demonstrations by an expert (represented as state/action pairs), as training data [15]. In end-to-end approaches, the state coincides with the robot sensing; this yields reactive controllers unless recurrent models are used to capture temporal dynamics [6].

The ground truth used for learning may have various origins: skilled drivers or operators [10], [14], [1], [3]; people walking [11], [12] or driving a vehicle that is not the target robot [13]; random controllers that sometimes lead to collisions, which the model learns to avoid [16]; hand-designed controllers [2]; controllers learned through reinforcement learning [9]; or the future position of vehicles in a large driving dataset [5]. Most of these approaches are stateless and reactive, but applications of Deep Recurrent Neural Networks allows to capture temporal dynamics in end-to-end approaches [6].

Compared to mediated approaches, end-to-end learning has been found in some cases to be slower to converge [17], [18] and to require more training samples [19]. Direct perception methods [20] are a further alternative for learning higher-level representations of the environment than mediated methods.

The task we consider in this paper is aimed to proximity human-robot interaction [21], [22]: a drone should be able to fly at an appropriate distance in front of people, following them [23] while waiting for possible command gestures [24]. The use of visual markers [25] simplifies this task but, in general, more sophisticated techniques, like Tracking-Learning-Detection [26], are needed [27]. For instance, several deep learning approaches have been proposed to estimate the 6D pose of a person's head [28] or body [29] from monocular cameras (as well of general objects [30]). Adopting such perception modules would be a reasonable alternative to solve our task; however, in this paper the task acts as a model of a larger class of tasks.

## II. TASK AND MODEL

We consider one specific task: controlling a quadrotor to stay at a fixed distance ( $\Delta = 1.5\text{m}$ ) and at eye-level height in front of a user who is free to move in an environment.

The available inputs are the video feed from a forward-pointing camera  $\mathbf{x}_{\text{im}}$  and the current linear velocity of the quadrotor  $\mathbf{x}_{\text{odom}} = (v_x, v_y)$ , obtained through ego-motion sensors; the controller outputs  $\mathbf{u}$  are composed by:

- the desired pitch and roll of the drone, which map to the acceleration along the drone's  $x$  and  $y$  axes ( $u_{ax}, u_{ay}$  respectively);
- the desired velocity along the  $z$  axis ( $u_{vz}$ );
- the desired angular velocity around the  $z$  axis ( $u_{\omega z}$ ).

The state  $\mathbf{s}$  for the given task can be compactly represented using two pieces of information  $\mathbf{s} = (\mathbf{s}_{\text{pose}}, \mathbf{s}_{\text{odom}})$ .

- A 3D transformation  $\mathbf{s}_{\text{pose}}$ , representing the pose (location and orientation) of the user's head with respect to the quadrotor; more specifically, only the heading component of the head's orientation is relevant to the task (in fact, tilting the head or looking up or down should not affect the drone behavior). Therefore, we represent with angle  $s_\theta$ ,  $-\pi \leq s_\theta \leq \pi$ , the relative angle between the user's and drone's orientation, such that  $s_\theta = 0$  if the user's face points along the negative  $x$  axis of the drone (i.e., the user is in front of the drone and faces towards it). Therefore,  $\mathbf{s}_{\text{pose}} = (s_x, s_y, s_z, s_\theta)$ .
- In addition, the state also contains the quadrotor's current measured velocity, which is available directly as an input  $\mathbf{s}_{\text{odom}} = \mathbf{x}_{\text{odom}}$ .

In this specific task, we can also design a controller  $C$  that produces control signals given the state:  $\mathbf{u} = f_C(\mathbf{s})$ , as detailed in Section III-B. Note that for other tasks (e.g. grasping), designing a controller might be very complex.

### A. General Approaches

We now describe the three approaches that we will compare in the following sections (see Figure 1). Even though for clarity we use the task-specific notation introduced above, these approaches can be easily generalized to other control tasks; then,  $\mathbf{x}_{\text{im}}$  corresponds to the subset of the inputs that we want to learn how to interpret in order to obtain high-level state information  $\mathbf{s}_{\text{pose}}$ .  $\mathbf{x}_{\text{odom}}$  represents inputs that we can (directly or through some processing step) turn into meaningful high-level state information  $\mathbf{s}_{\text{odom}}$ . For example, in an autonomous driving scenario:  $\mathbf{x}_{\text{im}}$  could correspond to lidar data;  $\mathbf{s}_{\text{pose}}$  to the relative position of other cars on the road and their velocities;  $\mathbf{s}_{\text{odom}}$  to the car odometry and to the car's position on the lane (obtained through some existing modules using inputs from  $\mathbf{x}_{\text{odom}}$ );  $\mathbf{u}$  to steering and acceleration/brake controls.

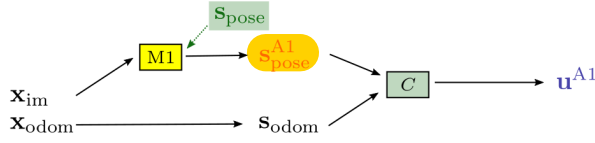
In **approach A1 (mediated)**, we learn a model  $M1$  mapping  $\mathbf{x}_{\text{im}}$  to  $\mathbf{s}_{\text{pose}}$ . Let  $f_{M1}$  denote the function implemented by  $M1$ ;  $\mathbf{s}_{\text{pose}}^{A1} = f_{M1}(\mathbf{x}_{\text{im}})$  denotes the estimate of  $\mathbf{s}_{\text{pose}}$ .  $\mathbf{s}_{\text{pose}}^{A1}$  is then joined to the odometry (available directly in the input) to form  $\mathbf{s}^{A1} = (\mathbf{s}_{\text{pose}}^{A1}, \mathbf{s}_{\text{odom}})$ . The control signals are then computed as  $\mathbf{u}^{A1} = f_C(\mathbf{s}^{A1})$ . In order to train  $M1$ , we require ground truth information on  $\mathbf{s}_{\text{pose}}$ . Moreover, this approach requires the availability of a controller  $C$ .

In **approach A2 (end-to-end)**, we directly learn a function  $f_{M2}$  to predict  $\mathbf{u}$  from  $\mathbf{x} = (\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{odom}})$ . We denote the result

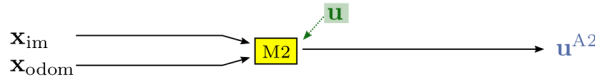
as  $\mathbf{u}^{A2} = f_{M2}(\mathbf{x})$ . Training M2 requires that a ground truth for  $\mathbf{u}$  is available. Such a ground truth can be obtained in two ways. *Learning strategy 1*: acquiring the ground truth for  $\mathbf{u}$  directly (e.g., recording a skilled human pilot); *Learning strategy 2*: obtaining the ground truth for  $\mathbf{u}$  through a designed controller  $C$ , which is given ground truth  $\mathbf{s}_{\text{pose}}$  information.

In **approach A3 (mediated with learned controller)** we learn a model M1 just like in approach A1. This model obtains a state estimate  $\mathbf{s}^{A3} = \mathbf{s}^{A1}$ . However, instead of using an hand-designed controller  $C$  to produce the control signals from the estimated state, we learn a mapping  $f_{M3}$  from  $\mathbf{s}$  to  $\mathbf{u}$ . The control signals are then computed as  $\mathbf{u}^{A3} = f_{M3}(\mathbf{s}^{A3})$ . In order to train M1, we require ground truth information on  $\mathbf{s}_{\text{pose}}$ . Moreover, in order to train M3 we need a ground truth on  $\mathbf{u}$ , e.g. from a human controller. Differently from A1, this approach does not require one to design the controller  $C$ .

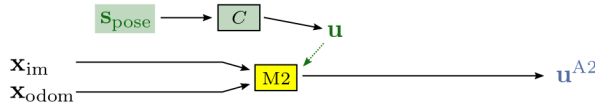
Approach A1 (mediated)



Approach A2 (end-to-end) training strategy 1



Approach A2 (end-to-end) training strategy 2



Approach A3 (mediated with learned controller)

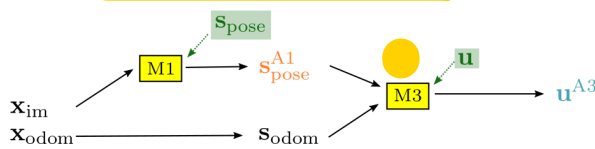


Fig. 1: Yellow background: learned models. Ground truth data used for training them is depicted with a dotted arrow. Requirements (in terms of ground truth or of a designed controller) are drawn with a green background.

### III. EXPERIMENTAL SETUP

#### A. Hardware and Infrastructure

All experiments take place in a  $10 \times 10$  m room fitted with an optical motion capture system. Within the room, a flying area of  $7 \times 7$  m is defined and fenced by a black lightweight net.

We use a Parrot Bebop 2 quadrotor; the quadrotor is networked via wifi to a server, and controlled through a ROS interface [31], which exposes at 5 Hz the velocity of the

drone  $\mathbf{x}_{\text{odom}}$  computed from the visual odometry of a bottom-looking camera and, at 30 Hz, the front-facing camera feed. The drone uses a sonar and an IMU to estimate altitude and attitude; an onboard low-level controller accepts inputs  $\mathbf{u}$  and updates the motors speed.

The quadrotor is outfitted with a mocap target, so that its exact 6D pose is acquired in real time. Moreover, we outfit the user's head with another mocap target; this is implemented by having the user wear one of two objects on their head, on top of which the mocap target is fixed: either a black baseball hat, or a thin elastic band. While the former is clearly visible when the user is seen from the front, the latter is almost invisible.

#### B. Controller

We implement a simple baseline controller  $C$  as a stateless function  $f_C$ . From the estimation of the person's pose, we compute a target point  $\bar{\mathbf{p}} = (s_x, s_y, s_z) + \Delta e(\pi + s_\theta)$  in front them, at distance  $\Delta$  along unit vector  $e(\pi + s_\theta)$ . Then, we compute a desired velocity  $\bar{\mathbf{v}} = [\frac{\bar{p}_x}{2\tau}, \frac{\bar{p}_y}{2\tau}]_{-v_{\max}}^{v_{\max}}$  to reach  $\bar{\mathbf{p}}$  in  $2\tau$  time, limiting the components magnitude to  $v_{\max}$ . Finally, we compute a control output to reach velocity  $\bar{\mathbf{v}}$  and rotate towards the user in  $\tau$  time:

$$\begin{aligned} (u_{ax}, u_{ay}) &= [((\bar{v}_x, \bar{v}_y) - (v_x, v_y)) / \tau]_{-a_{\max}}^{a_{\max}} \\ u_{vz} &= \bar{v}_z \\ u_{\omega z} &= [\theta(\bar{\mathbf{p}}) / \tau]_{-\omega_{\max}}^{\omega_{\max}}, \end{aligned}$$

where  $\theta(\bar{\mathbf{p}})$  is the azimuth of vector  $\bar{\mathbf{p}}$ . Parameters are fixed to  $\tau = 0.5$  s,  $v_{\max} = 1.5$  m s<sup>-1</sup>,  $a_{\max} = 1.0$  m s<sup>-2</sup>,  $\omega_{\max} = 2.0$  rad s<sup>-1</sup>.

#### C. Dataset Acquisition

In order to ease the acquisition of the dataset, the drone follows a controller  $C_{\text{acq}}$  that, given the pose of the user and the pose of the quadrotor provided by the mocap system, generates control signals that keep the quadrotor at a parametrized distance (that we adjusted between 1.0 and 2.0 meters during the acquisitions) in front of the user's head. Such a controller is similar of the one described in Section III-B. The outputs of  $C_{\text{acq}}$  are not part of the dataset: they are only used to facilitate the acquisition of the sessions and make for an engaging experience for our test subjects.

We recorded 15 different sessions, each with a different user (age 23 to 38, different ethnicity, height ranging from 160 to 197 cm, variable levels of physical fitness, different clothing styles, hairstyles and colors). During the acquisition, users are instructed to move around the room freely: the quadrotor will then fly to stay in front of their head. After an initial period with cautious motion, users start challenging the controller and moving rather aggressively (videos of recording sessions are available as supplementary material), so that the quadrotor struggles to keep up; this ensures that many different relative positions for the head and the quadrotor are represented in the data. In some recordings, the users wear the hat, while in others the headband (which is almost not visible in the camera frames). The room is equipped with



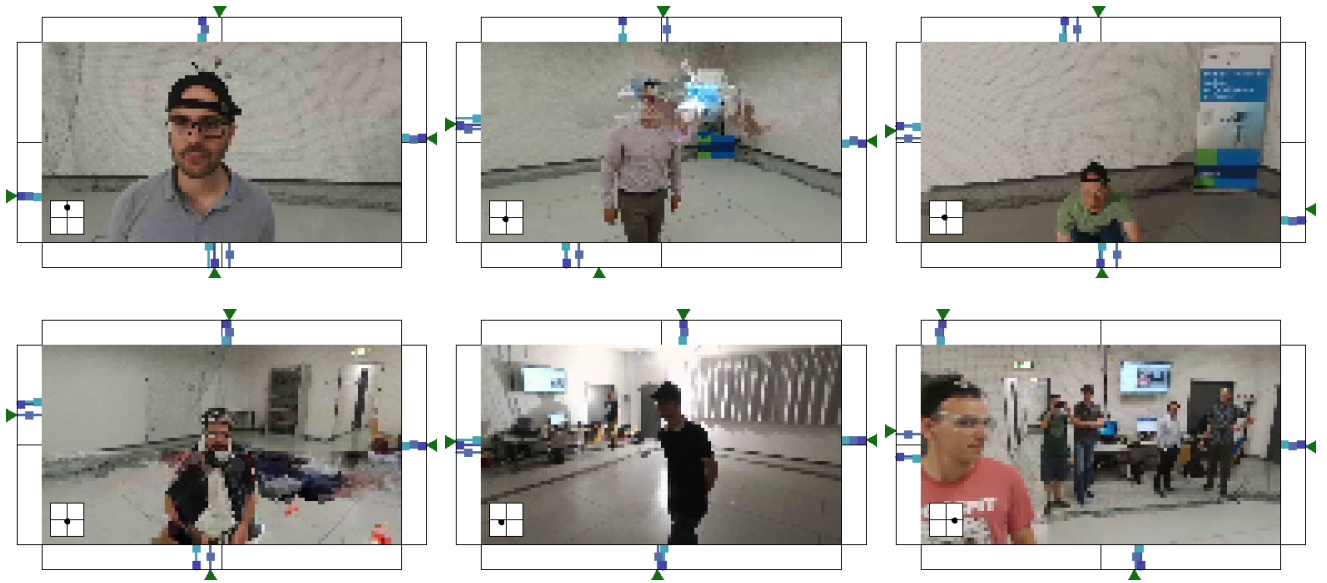


Fig. 2: Instances from the dataset. Each shows ground truth (green triangle) and outputs (blue rectangles) of A1, A2, A3 for each control variable:  $u_{ax}$  (left),  $u_{ay}$  (bottom),  $u_{vz}$  (right),  $u_{\omega z}$  (top). More details and videos in supplementary material. Note that some images in the dataset are affected by acquisition, compression or transmission artifacts.

4 distinct light sources (2 overhead, 2 movable spotlights placed in different corners of the room), which we toggle and move at different times of the recording to add variability to the scene. During some of the recordings, more people are in the flying area; moreover, often the quadrotor camera also sees people and different background objects standing outside of the flying area, such as computers, screens, desks, and windows. While acquiring the data, we take care that, if more than one person is visible in the frame, the person wearing the mocap target is the one closest to the camera. On occasion, the video link is temporarily corrupted and some frames are acquired with visual artifacts; we purposefully do not remove such frames from the recording as most of them can be still understood by an human observer. On average each session is 3 minutes long, totaling more than 45 minutes of flight.

All sessions are recorded in ROS bagfiles, from which we extract 79k dataset instances at a rate of 30Hz. Each instance contains: the video frame  $\mathbf{x}_{im}$ , the pose of the user head relative to the drone  $\mathbf{s}_{pose}$ , and the drone velocity  $\mathbf{x}_{odom} = \mathbf{s}_{odom} = (v_x, v_y)$ . The controller  $C$  is then applied to  $(\mathbf{s}_{pose}, \mathbf{s}_{odom})$  to obtain the ground truth value  $\mathbf{u}$  (containing the four control variables). Note that the ground truth control signal does not necessarily correspond to the control signal that the drone was receiving during the acquisition, which came from a different controller.

We use all data from three sessions (16k instances) as a test set to quantitatively evaluate the performance of the three approaches. We randomly split the 63k instances from the remaining 12 sessions into a training set (50k instances) and a validation set (13k instances).

#### D. Machine Learning Models

1) *Model M1*: Model M1 is implemented as a deep neural network using a ResNet [32] architecture, which has been recently adopted for similar quadrotor control tasks [13].

The network accepts as input a  $108 \times 60 \times 3$  RGB image, and produces the 4 components of  $\mathbf{s}$ . The internal architecture is a ResNet-8 [32] followed by two dense layers with 256 and 128 neurons, respectively.

2) *Model M2*: Model M2 has a similar architecture as M1, but additionally accepts  $\mathbf{x}_{odom}$  through two additional input neurons that skip the ResNet-8 layers. These neurons, concatenated to the ResNet-8 output, are input to the two dense layers with 256 and 128 neurons. The outputs correspond to the 4 components of  $\mathbf{u}$ .

3) *Model M3*: Model M3 is implemented through a simple multilayer perceptron that maps 6 input values ( $\mathbf{s}^{A3} = \mathbf{s}^{A1} = (\mathbf{s}_{pose}^{A1}, \mathbf{s}_{odom})$ ) to the 4 components of  $\mathbf{u}$ ; it contains 2 hidden layers, with 256 and 128 neurons respectively.

4) *Training*: We train the three models with the same setup: 1) we use Mean Absolute Error as loss function and ADAM [33] as optimizer with a learning rate of 0.001; 2) we speed up learning by reducing the learning rate when the validation loss plateaus for more than 5 epochs; 3) we use early stopping (10 epochs of patience on the validation loss, with a maximum of 200 epochs).

#### IV. EXPERIMENTAL RESULTS

We report three sets of experiments. First, we quantitatively evaluate the prediction quality of different approaches for all instances in the testing set, versus the ground truth for  $\mathbf{u}$ , also evaluating the impact of the training set size; second, we compare the trajectories of a quadrotor when

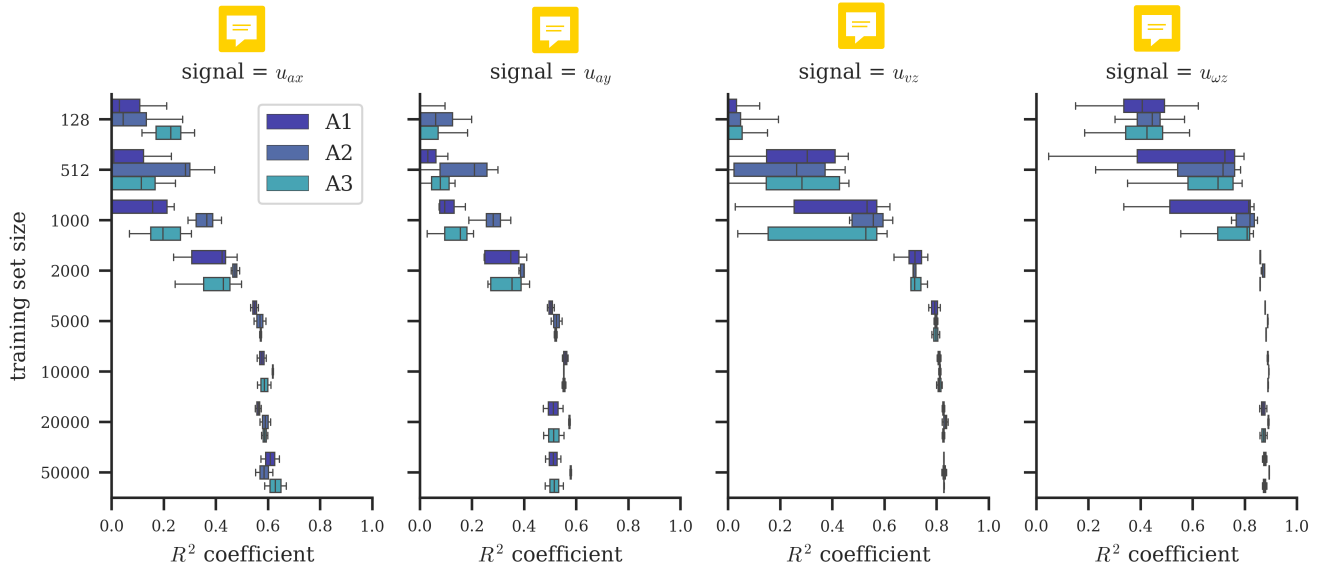


Fig. 3: Quantitative results for the four control variables  $\mathbf{u}$ : estimation quality versus number of training samples.

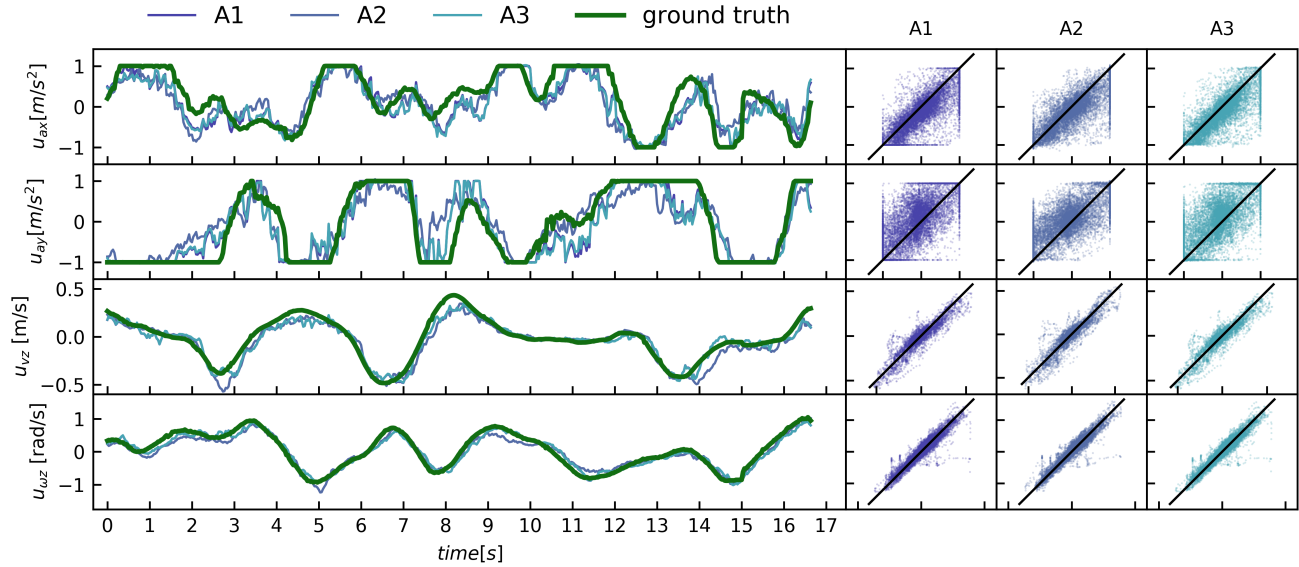


Fig. 4: We report each of the four control variables on a separate row. Left: the time series of the ground truth control (green thick line) and of the outputs of the ML approaches (thin lines), for a 16 second segment extracted from the test set, in which the user was moving very aggressively. Right: the scatter plot comparing, on the y axis, the output of A1 (left), A2 (center), and A3 (right) to the ground truth (x axis), for the whole testing set; a perfect model would yield all points on the diagonal.

it is controlled by an ideal controller, with the trajectories obtained by A1, A2 and A3. Finally, we perform qualitative robustness tests on the system controlled by each of the three approaches.

#### A. Quantitative results on testing instances

For each component of  $\mathbf{u}^{A1}$ ,  $\mathbf{u}^{A2}$  and  $\mathbf{u}^{A3}$  we compute the coefficient of determination  $R^2$  [34] of the estimate. This measure corresponds to the proportion of variance in  $\mathbf{u}$  that is explained by the model: a perfect estimator yields  $R^2 =$

1; a dummy estimator that always returns the mean of the variable to be estimated yields  $R^2 = 0$ ; even worse estimators yield  $R^2 < 0$ .  $R^2$  allows us to compare the quality of our estimate for different components of  $\mathbf{u}$  even though each has a different variance.

We compute these metrics for the three approaches trained on different amounts of training data; in particular, we are interested to compare how hard it is for the different approaches to achieve a given performance. For a given training set size  $T$ , we randomly sample  $T$  out of the 50k

training instances, which we use to train the models for each of A1, A2 and A3. Validation and testing datasets in all cases remain the same. To account for the variability due to the sampling of the training set (which is very large for low values of  $T$ ), we repeat each experiment for up to 50 replicas.

Figure 3 reports the  $R^2$  values resulting from the experiments described above. We observe that prediction quality increases with  $T$ , but plateaus after  $T = 10000$ ; there is no clear difference in prediction performance among A1, A2 and A3.

$u_{\omega_z}$  and  $u_{v_z}$  yield  $R^2$  coefficients of 0.88 and 0.82, respectively. Prediction quality is significantly worse for  $u_{a_x}$  and especially for  $u_{a_y}$ , with  $R^2$  values of 0.59 and 0.57, respectively. This is easily explained:  $u_{\omega_z}$  is mostly dependent on the position of the body in the frame, which is easy to perceive in  $\mathbf{x}_{im}$ ;  $u_{\omega_z}$  is learned significantly better than a dummy regressor, with as few as 128 training instances.  $u_{v_z}$  mostly depends on the vertical position of the head in the frame, which is also easy to perceive in  $\mathbf{x}_{im}$ , but has less variability in the datasets. Predicting  $u_{a_x}$  and  $u_{a_y}$  is harder: the former relies on an accurate perception of the distance of the user (which is confounded by their height and body size), and the latter on an estimate of the relative orientation of the head ( $s_\theta$ ), which is arguably hard to get on low resolution inputs.

Note that prediction quality is not necessarily related to the quality of the resulting robot behavior. For example, a model which always yields a very tiny but systematic overestimate on  $u_{a_x}$  yields an  $R^2$  value close to 1, but would cause the quadrotor to crash on the user in a very short time. On the contrary, a model that produces predictions affected by large amounts of uncorrelated noise with zero mean could yield acceptable behaviors (especially with short control timesteps) but very disappointing quantitative metrics. To provide a better idea of the usability of the predictions for control, Figure 4 illustrates the outputs of the three approaches (trained with  $T = 50000$ ) on a short segment extracted from a recording belonging to the testing set. We observe that the predicted signals closely track the ground truth; outputs for  $u_{a_x}$  and  $u_{a_y}$  exhibit a larger amount of high-frequency noise, which appears non-systematic.

### B. Analysis of flying performance

Figure 5 compares the trajectories resulting from the ground truth controller, with those resulting from A1, A2, and A3. The quadrotor approaches a user (not in the dataset), who is standing still, from different relative poses: the final pose of the robot always faces the user, and is reached in about 5 seconds, after which the robot stabilizes in a short time.

In Figure 6, the quadrotor’s approach towards a different user (also not in the dataset) was run 5 times for A1 and A2. This is the only case in our tests where we could notice a (small) difference in the behavior of different approaches, with A2 being somewhat smoother and closer to the ground truth trajectory.

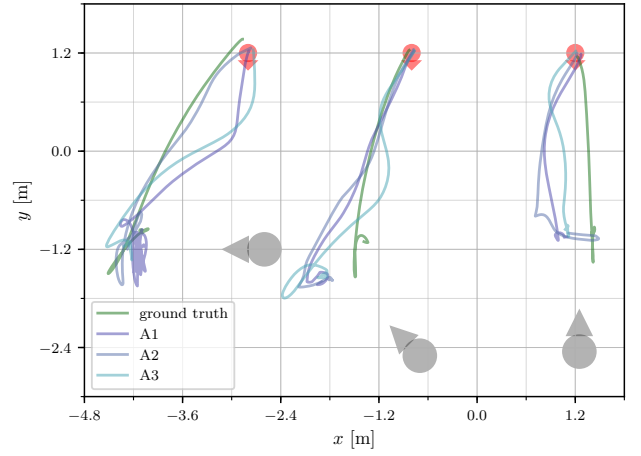


Fig. 5: Trajectories followed by the quadrotor starting in the red pose while approaching a person standing still in different poses (black circles), from left to right: initially facing the quadrotor at  $90^\circ$ ,  $45^\circ$ , and  $0^\circ$ . Trajectories represent one run and are colored by the controller in use.

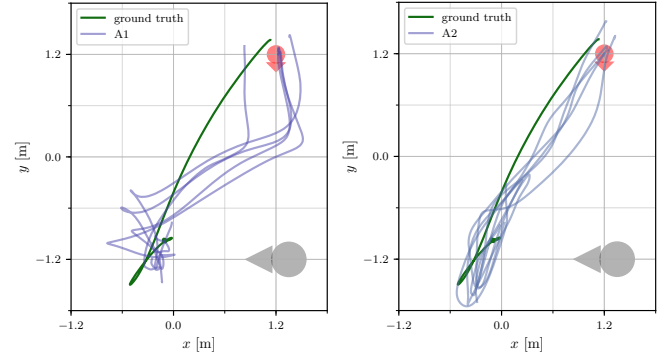


Fig. 6: Five runs for each controller (left: A1, right: A2) towards a person initially facing the quadrotor at  $90^\circ$  and a single run with the ground truth controller (green).

Supplementary videos report more details on this experiment and extensive tests where we challenge the system robustness by having multiple people in the frame, quick movements, distracting objects, sudden and extreme lighting variations. During the tests, we cycled the control between the three approaches, and found that they behaved indistinguishably from each other; in all cases the drone behavior was predictable and perceived as safe by the users.

## V. CONCLUSIONS

We considered the task of controlling a quadrotor to hover in front of a freely moving user using input data from an onboard camera. On this task, we compared mediated approaches to end-to-end approaches; we found equivalent quantitative performance, learning difficulty, perceived quality of the robot behaviors and robustness to challenging inputs; in only one occasion we measured repeatably different trajectories, which were slightly smoother in the end-to-end approach, but correct in all cases and perceived as very similar.

## REFERENCES

- [1] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Advances in neural information processing systems*, 2006, pp. 739–746.
- [2] E. Kaufmann, A. Loquercio, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, "Deep drone racing: Learning agile flight in dynamic environments," in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87, 2018, pp. 133–145.
- [3] M. Bajracharya, A. Howard, L. H. Matthies, B. Tang, and M. Turmon, "Autonomous off-road navigation with end-to-end learning for the lagr program," *Journal of Field Robotics*, vol. 26, no. 1, pp. 3–25, 2009.
- [4] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1856–1860.
- [5] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3530–3538, 2017.
- [6] L. Chi and Y. Mu, "Deep steering: Learning end-to-end driving model from spatial and temporal visual cues," *arXiv preprint arXiv:1708.03798*, 2017.
- [7] J. Heylen, S. Iven, B. D. Brabandere, J. O. M., L. V. Gool, and T. Tuytelaars, "From pixels to actions: Learning to drive a car with deep neural networks," in *2018 IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 606–615.
- [8] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.
- [9] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.
- [10] S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert, "Learning monocular reactive uav control in cluttered natural environments," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 1765–1772.
- [11] A. Giusti, J. Guzzi, D. C. Ciresan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.
- [12] N. Smolyanskiy, A. Kamenev, J. Smith, and S. 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*, 2017, pp. 4241–4247.
- [13] A. Loquercio, A. I. Maqueda, C. R. del Blanco, and D. Scaramuzza, "Dronet: Learning to fly by driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1088–1095, 2018.
- [14] S. Kumaar, A. Sangotra, S. Kumar, M. Gupta, S. Omkar *et al.*, "Learning to navigate autonomously in outdoor environments : Mavnet," *arXiv preprint arXiv:1809.00396*, 2018.
- [15] A. Attia and S. Dayan, "Global overview of imitation learning," *arXiv preprint arXiv:1801.06503*, 2018.
- [16] D. Gandhi, L. Pinto, and A. Gupta, "Learning to fly by crashing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 3948–3955.
- [17] T. Glasmachers, "Limits of end-to-end learning," in *Proceedings of the Ninth Asian Conference on Machine Learning*, vol. 77. PMLR, 2017, pp. 17–32.
- [18] S. Shalev-Shwartz, O. Shamir, and S. Shammah, "Failures of gradient-based deep learning," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 3067–3075.
- [19] S. Shalev-Shwartz and A. Shashua, "On the sample complexity of end-to-end training vs. semantic abstraction training," *arXiv preprint arXiv:1604.06915*, 2016.
- [20] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*. IEEE Computer Society, 2015, pp. 2722–2730.
- [21] W. S. Ng and E. Sharlin, "Collocated interaction with flying robots," in *2011 RO-MAN*, 2011, pp. 143–149.
- [22] T. Naseer, J. Sturm, and D. Cremers, "Followme: Person following and gesture recognition with a quadcopter," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 624–630.
- [23] M. J. Islam, J. Hong, and J. Sattar, "Person following by autonomous robots: A categorical overview," *arXiv preprint arXiv:1803.08202*, 2018.
- [24] E. Peshkova, M. Hitz, and B. Kaufmann, "Natural interaction techniques for an unmanned aerial vehicle system," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 34–42, 2017.
- [25] F. Vasconcelos and N. Vasconcelos, "Person-following uavs," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.
- [26] Z. Kalal, K. Mikolajczyk, J. Matas *et al.*, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.
- [27] R. Barták and A. Vykovský, "Any object tracking and following by a flying drone," in *2015 Fourteenth Mexican International Conference on Artificial Intelligence*, 2015, pp. 35–41.
- [28] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognition*, vol. 71, pp. 132 – 143, 2017.
- [29] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, 2016, pp. 561–578.
- [30] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.
- [31] M. Monajjemi, "Bebop autonomy," <http://bebop-autonomy.readthedocs.io>, 2015.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [34] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014, vol. 326.