**CRAIG:** Hi, I’m Craig Smith and this is Eye on AI.

This week, I talk to Vladlen Koltun, who has run research teams at Intel and now Apple, though he is here as a guest on the podcast in his personal capacity. Vladlen has worked on the coupling of locomotion and artificial intelligence, working with four-legged robots and drones, and on the creation of a high-throughput computer vision training environment that works at an incredible million-frames per second.

The work itself is fascinating, but what is most interesting is how that works fits into a theory of intelligence that sees higher-level intelligence emerging from this more basic coupling.

But before we begin, I want to give a shout out to my sponsor, [ClearML](https://t.clear.ml/eye_on_ai), an open-source MLOps solution. Users can customize the solution, which allows you to easily Track, Orchestrate, and Automate ML Workflows at Scale.

Whether you're a Data Engineer, ML engineer, DevOps, or a Data Scientist, [ClearML](https://t.clear.ml/eye_on_ai) is the best collaborative MLOps tool on the market, with full visibility and extensibility.

Setup your own server or Signup for free, log your entire process, version data, build model repositories, provision machines, schedule containers, and deploy pipelines directly from code.

Do me a favor and checkout ClearML – at [clear.ml](https://t.clear.ml/eye_on_ai)! They are supporting this podcast, so if you’re a listener, let’s support them.

Meanwhile, I hope you find the conversation as amazing as I did.

**CRAIG:** So, go ahead, Vladlen, introduce yourself to our listeners.

**VLADLEN:** Sure. I'm a researcher. I work in robotics, computer vision, machine learning, computer graphics. Sometimes I've entered into other areas. I've been doing this for a while, a couple of decades. I've run research organizations in a number of institutions.

**VLADLEN:** Currently I'm at Apple and I'm building a research organization at Apple. Again, I'm not representing my employer here. I'm doing this in a personal capacity. Before I was the chief scientist for intelligence systems at Intel, where I built up an international research lab active in robotics, machine learning, computer vision, and so forth.

**VLADLEN:** And this is what I do.

**CRAIG:** Yeah. I'm curious, I'm always curious about people's journeys. How did you get involved in machine learning? Where did you go to school?

**VLADLEN:** Yes, I did my PhD in Tel-Aviv University. So after we left the Soviet Union, we moved to Israel and I lived there for nine years. And that is where I completed my higher education.

**VLADLEN:** I got my PhD in computer science. I was a theoretician. I got my PhD in theoretical computer science. And after that, I moved to UC Berkeley as a post-doc in the theory group. I did my postdoc with Christos Papadimitriou, a well-known theoretical computer scientist. And then I moved to Stanford as a faculty member, as a professor in theory.

**VLADLEN:** I was hired at Stanford as a, as a theoretician. So I moved into, into applied work after that. I moved into computer graphics and from there expanded to computer vision and then machine learning and then robotics. So it has been a gradual evolution, a gradual expansion of research interest.

**CRAIG:** One of the things I'm interested in with all the researchers I speak to, but particularly you, how all of the work fits together. Your recently published works involves robotic perception, including drones and four legged robots. Can you tell us about how you came to focus on that area and put it into the context of what's happening in ML and robotics generally?

**VLADLEN:** Yeah. Well, you know, sometimes when we explain what we do, in effect, we fit stories after the fact to things that happen. So probably what actually happened on the ground is that I got interested in something, it's even hard to explain originally why, but something attracted my attention.

**VLADLEN:** So I started doing it, then something else attracted my attention. And I started doing that. And a couple of hundred projects later, we can look back and fit a story, and the story rings true. I can tell you the story that rings most true to me, but the disclaimer is that we never really know, because it is all stories that we fit after the facts, the story that that does ring true, and it has been quite consistent for years now for me, is that there are two questions that I just find intrinsically interesting, I just get really excited about, I want to think about them. I want to make progress in anything that can shed light on these questions. It fires me up and I want to work on it.

**VLADLEN:** So when I see some technique that is relevant to that, that can help me make progress, I want to understand it. I want to use it. I want to extend it. And question number one is the question of photo realism, of presence. So creating virtual realities that feel real. Creating virtual worlds that look real and feel real and, to some extent, allow us to transcend our physical location and our physical bodies. And that is actually what originally attracted me to computer science before I was a theoretician. And before I got my PhD in theory, I taught myself how to program in order to do computer graphics. I saw computer graphics and my teams, and it was a love at first sight.

**VLADLEN:** I just, I just wanted to do it. It looked magical. It seemed magical. And this feeling of magic never really went away. It's still feels magical to me and I still wanted to do it. And a lot of my work still has to do with understanding for the realism, understanding presence, understanding, simulation, understanding how to build virtual worlds that to some extent, reproduce aspects of reality.

**CRAIG:** That interest in computer, was that, driven at all by an interest in video games? I know a lot of the researchers I spoke to started in computer graphics because of an interest in video games.

**VLADLEN:** Interestingly enough for me, it was, it was different. I did play video games. I did enjoy, video games, but what really gripped me is when I saw computer graphics being used as an artistic medium.

**VLADLEN:** So, at the time there was a kind of artistic scene called the Demoscene, it might still be active to some extent, where the idea was that groups of people, groups of enthusiasts came together and created almost like pieces of art with computer graphics that blended original digital art with music and a typical production would be a few minutes of visual effects and it was real-time computer graphics.

**VLADLEN:** And the idea was to blend technology and art. You needed to be an extremely good programmer, who would optimize real time computer graphics to produce very impressive, visual effects, but also blend in art and blend in, music. And there was a large number of people all around the world involved in this community. And people came together for competitions. And, I was fortunate enough, completely by chance, by coincidence, to meet some people in Israel who were active in this community. And they showed me their work and it was just incredible. So computer games to me didn't really motivate me at the time to learn to program so that I could build computer games, computer games.

**VLADLEN:** I was actually fine just consuming them and playing, but I wasn't really driven to create computer games. Um, but with, with the demo scene, when I, when I saw these beautiful, pieces of creative expression, I just wanted to understand the medium and I wanted to participate in that. So that's how it started.

**CRAIG:** Yeah, that's fascinating. And then you moved into, into robotics and combining, perception with robotics. I wanted to ask you about your most recent paper on, on, four legged robot controllers.

**VLADLEN:** Yeah, so to make the transition smoother, um, led me mention the other big question. The second big question that has, that has driven my work. And that is the question of intelligence understanding and intelligence, reproducing intelligence in physical systems. That's the other question that when it hit me, I just wanted to understand it and do whatever I can to understand it better through my work.

**VLADLEN:** And there are different approaches to understanding intelligence, but one approach is to look at nature and to look at evolution and to look at how intelligence systems came to be in the physical world. And if you look at nature, and if you look at what evolution has spent its time doing, it spent hundreds of millions of years, optimizing perception and action, and the coupling of perception and action.

**VLADLEN:** If you look at animals they perceive in order to act, and they act in order to perceive and perception and action interact in this closed loop, that gives rise to what we see as intelligent behavior. And animals maintain certain internal representations, representations that mediate perception and action.

**VLADLEN:** That inform perception, that are informed by perception. And that inform action, are informed by the consequences of actions. And I want to understand that. And the hypothesis is that if we really thoroughly understand this embodied intelligence at a level that we can see in animals, it doesn't even have to be an agent that can play chess and compose symphonies.

**VLADLEN:** Okay. The level of animal, if you look at what animals are doing, if we understand that, we will be well on our way towards understanding higher forms of cognition and intelligence in the most general sense, because again, looking back at evolution, it seems that higher level cognition and all the chess playing and symphony composing were built on top of this massive foundation of embodied intelligence and functioning in physical environments and coupling perception and action to perverse and interact with physical environments to accomplish very physical embodied tasks. And in fact, if you look at the how long it took evolution to build higher level cognition on top of this foundation, it happened very fast.

**VLADLEN:** The foundation of physical intelligence, of embodied intelligence to kindreds of millions of years. All the chess playing and symphony compositing seemed much easier, at least if you look at how long it took evolution to derive these skills. So what's driven my work in robotics is essentially a program of trying to reproduce the essential skills, the essential functionality we see in animals, in artificial systems. Because they have offices, if we can reproduce the level of agility, adaptability, or robustness, generality of, of animals in physical systems, then we will understand a lot about intelligence and we will probably understand something essential that we can build the rest on, on top of.

**VLADLEN:** So we have a line of work with amazing collaborators on a quadrupedal robots, legged robots that traverse natural environments that are incredibly robust, they can go anywhere. Rain, snow, mud, hills, rough terrain, hiking, anything you want. And they keep going. And they're incredibly robust.

**VLADLEN:** We have a line of work on drones on flying machines that can at high speed traverse natural environments like fly through a dense forest at high speed, avoid trees, and so forth. Likewise requires very tight coupling between perception and action and perception. That informs action in a very tight loop.

**VLADLEN:** The work that you asked about the robust, legged locomotion that integrates visual perception and proprioception is part of that program. It's a step in pushing the adaptability and generality and the robustness of embodied systems that we can deploy in the physical world. To understand the step that was taken in this particular work, we can look at the previous step where we deployed a locomotion controller that was essentially blind. It's perception was just the perception of the robots body.

**VLADLEN:** The robot’s configuration. It's acceleration. It's angular, acceleration the configuration of the limbs, the joints, the joint angles, and how the joint angles are changing through time. And just by perceiving the state of its own body, we showed that the robot could keep itself stable, keep itself up upright and locomote through the environment.

**VLADLEN:** And in fact, do it better, more robustly than prior systems. Now, if you're blind, there are certain things that you just cannot do very well. You cannot traverse the environment at very high speed because you can't anticipate what's going to happen. You're essentially feeling out the environment with your own body.

**VLADLEN:** So the controller, the blind controller we trained in our previous war was quite cautious. It was quite methodical and cautious and it would walk and it would basically never fall, but it was quite slow because it needed to feel out the environment with its body. In the current work, we added visual perception and we showed that we can do it in a very flexible, adaptive way.

**VLADLEN:** So that even when visual perception is misleading or it misfires, somehow the robot still never fails. It never falls. But it can also leverage visual perception to traverse the environment much, much faster.

**CRAIG:** That's interesting. I didn't realize or understand from the paper. I mean, I haven't followed your previous work on this, that, that you were adding visual perception. I thought it was the other way around, frankly. Just a question about, about the evolution of higher level intelligence from this coupling of physical action and control, two questions. One is, do you think higher level intelligence then emerges because the networks become so complex for motor control, that the hardware is in place for something higher level to emerge?

And, your work, for example, on the quadrupedal robots, in your mind, are you tackling these problems as part of this larger effort to understand how systems like this develop, or is it really applied research where, how can we make four legged robots more perceptive and mobile?

**VLADLEN:** Yes. Let me address these one at a time. Let's start with the emergence of higher level intelligence. I do genuinely believe that higher level intelligence emerged from the foundations of embodied intelligence and from the necessity of functioning in a more and more complex way in physical environments.

**VLADLEN:** So if you look at animals that need to traverse physical environments, they need to do it efficiently. They need to survive. They need to minimize physical damage. They need to obviously not fall, but also not get into various traps. So, they need to not get entrapped by some quick sand or some dangerous very rapid, whitewater stream, or they need to not fall off a cliff.

**VLADLEN:** They also need to, find sources of energy. They need to find food and as animals evolve in complexity, they actually start hunting each other. And that really ramps up the cognitive load, because they need to track other animals that can serve as sources of energy. And, as all animals become more and more intelligent it produces a kind of ratchet, a kind of feedback loop where as the prey gets more intelligent and adopt certain evasive maneuvers and more and more complex motion patterns, the predator needs to now have a mental model, a theory of mind of sorts of the prey. They need to predict and anticipate the preys actions and vice versa.

**VLADLEN:** Prey needs to have a mental model at theory of mind, a predictive model of the predator. And then we get cooperation. When cooperation happens, you need to have certain internal models, predictive models of other animals that you are cooperating with. Perhaps a pack of wolves cooperating and taking down a bison that no single wolf would be able to take down, but will provide weeks of nourishment.

**VLADLEN:** In this process, you learn to estimate and track the layout of the environment. You need to have a persistent, theoretical mental model of how the environment is arranged. You need to understand that there are objects, things that move coherently in the environment, and some of them are passive and. in accordance with some physical principles, like a log floating passively on a stream, and you will actually build mental model of these physical principles.

**VLADLEN:** Implicitly, you will not formulate Newton's laws and fluid dynamics. You will not as a wolf write down equations, but you will have to develop certain mental models, certain predictive models, in light of these principles. And you will learn that other objects move out of what seems like their own volition.

**VLADLEN:** And in fact, some of them react to you and you will start to construct these mental models of have they seen the me already? What are they going to do when they see me and so on and so forth? So all of this happens basically at a preLink stick level. If you look at wolves, if you look at cheetahs they're at that level.

**VLADLEN:** Okay. They're not going to write a novel, but, but they have this understanding of the world. Once you're at that level, once you have this understanding of the word. It seems like a fairly small step to start naming things to go, okay, this thing, this thing is called gazelle.

**VLADLEN:** Okay. This thing that you've been chasing your whole life, that gazelle. Okay. And this unpleasant feeling that, that you have that drives you to seek out gazelles increasingly more desperately, let's call that hunger. Okay. So that actually seems like a fairly small step now. Yes. There's grammar and positionality and increasingly abstract and self-referential concepts.

**VLADLEN:** Okay. We don't fully understand that. And we do not fully understand how that emerges from the physical layer. But it does seem that getting to kind of wolf, cheetah level will get us much, much closer to where we want it to be in terms of understanding intelligence as a whole.

**CRAIG:** In your research, you started talking about constructing narratives after the fact to fit a series of events. You were in a theory previously. Have you thought through this process of evolution of intelligence and then picked research projects that will inform that theory, or is this theory of the evolution of intelligence something that you can see your research fits into after the fact? I'm curious, how your thoughts evolve in that way.

**VLADLEN:** So this approach to artificial intelligence made sense to me and a spoke to me. It resonated with me from the very beginning. And I was exposed to this world view before I became active in machine learning and robotics. I generally always liked to go back to primary literature and trace the evolution of various fields to their foundations.

**VLADLEN:** And so before I became particularly active as a researcher in robotics and machine learning, I read older materials by many of the pioneers in the field. And in particular, there were researchers such as Hans Moravec and Rod Brooks that wrote eloquently about the embodied approach to intelligence and why it made sense, in the eighties, for example. And they were reacting against the logical schools of artificial intelligence that was dominant since the fifties.

**VLADLEN:** And I read the arguments of the logical school and I read the arguments of Hans Moravec and Rod Brooks and, the Moravec-Brooks arguments made much more sense to me. I would just read them side by side and it's kind of obvious that one of the sides, just makes much more sense, they just seem more aligned with reality.

**VLADLEN:** And to be that, that was definitely the more of a group, school pen that they didn't get everything. Right. But in essence, The argument of, looking at evolution, looking at the natural world, looking at the only existence proofs that we have of, the kind of intelligence that we want to create and looking at how it came about, it made sense.

**VLADLEN:** It it's true that we have existence proofs, and we can trace how these existence proofs came to be. We don't have to slavishly reproduce everything. We didn't get airplanesby building birds down to the feathers. But, we can look at the essential principles and embodiment does seem very helpful.

**VLADLEN:** And, it is quite sensible to see that if we get embodied intelligence, the rest should be clearer.

**CRAIG:** And, so your work, for example, on combining proprioception and exteroception, you’re probing steps in that evolution to inform your own view of how this happens or is it that somebody wants to optimize the quadrupedal robot and says, Vladlen, can you make this work better?And your world view about the emergence of intelligence informs that research. Which comes first.

**VLADLEN:** I think it might be a bit different for my collaborators than I. So for example, the line of work on legged locomotion is in collaboration with an amazing lab at ETH, ETH Zurich.

**VLADLEN:** And they're doing all the heavy lifting. They have their robots, they build their own. So they are doing everything that actually looks like work. I occasionally think and listen and speak, but they do everything that actually, actually has real work.

**VLADLEN:** And they come very strongly from a mechanical engineering background. And I think for some of them, the answer would be, we want to build beautiful legged machines. and we want the applications that legged machines will unlock. And, I want that, too. I also want beautiful, graceful, effective legged machines, and I think they will enable really, really fantastic, useful applications that augment human capability.

**VLADLEN:** But at the very core, I am primarily motivated by the scientific questions. I really want to understand intelligence and I just want to push on our understanding of physical embodied intelligence and the most effective way I know to push on that is by creating increasingly capable physical systems.

**VLADLEN:** Because if you're not actually creating actual functioning, physical systems, it's very easy to delude yourself. It's very easy to start theorizing and saying, well, this makes sense. That makes sense. And there's been just a bit too much of that in AI. You can see the first 30 years of AI as being maybe much too much on the plausible arguments side and not enough on let's get things done and let's create actual, capable, useful intelligence systems.

**VLADLEN:** So I feel like I understand things much better and I'm much more confident in my understanding if I participate in actually creating highly capable, physical systems.

**CRAIG:** On that work on the four legged robot, you said in one of the papers, I think your most recent paper, that the core of it is a recurrent encoder. And I understand what an encoder is. I understand what recurrent means. But, can you talk about how a recurrent encoder combines proprioception and an exteroception into what you called an integrated belief state? And is that something that's happening, continuously? Is there a clock that it's resetting, you know, as it would in a CPU?

**VLADLEN:** Yes, it is a continuously acting encoder. The encoder is basically a little neural network that gets as input a representation of the visual perception, the exteroception about the layout of the environment around it. And the proprioceptive state comes from sensors, such as inertial measurement units, and joint encodersthat give you information about the global movement, the global acceleration of the body of the robot through space and the angles and, accelerations of, of the various joints off of the robot. So it's a kind of sense of how your body is moving and how your limbs and joints are moving in in relation to each other.

**VLADLEN:** This proprioceptive state is generally good. It's generally true. If those sensors tell you that you're tilting and, and hurt the links through, through space, it's generally. The visual perception is provided by, by a depth sensor, um, LIDAR in this case. And these depth sensors have certain failure modes.

**VLADLEN:** So, for example, if you are seeing thick, dense vegetation, okay. then the depth sensor is going to return depth readings from kind of the top of that, of that vegetation. So it's going to look maybe like there is a mound ahead of you. But when you start walking through it, you will expect that you'll walk up onto the mound, but what's going to happen instead is that you'll start walking through it.

**VLADLEN:** The vegetation cannot support you because it's not solid. That is something that may not be apparent from just the geometric readings of, of the depth sensor. What recurrent encoder learns to do is it learns to correlate the information from visual perception and proprioception and detect very quickly, very adaptively when this information is out of whack.

**VLADLEN:** Okay. So the visual perception tells you, you should be on a solid surface and it should be supporting you at this height. And, the proprioception tells you actually you're tilting sideways and your way through that surface. And you're at a different height. You step through the thing that you thought is going to support.

**VLADLEN:** The encoder learns to start discounting the visual perception and passing a representation of, of the environment that is much, much closer to what proprioception says is true. So adaptively, it learns to take advantage and incorporate the visual perception information when it's consistent with proprioception.

**VLADLEN:** But when the two get out of alignment, it learns to very quickly adaptively and appropriately discount, the visual perception information. And it's important that it's done without some heuristics and thresholds. It's not like we said, well, when this parameter is sufficiently different from that parameter, and it's above that threshold don’t pay attention to that parameter.

**VLADLEN:** No, that's too much. We're not smart enough to do it. Nobody is. Nobody has succeeded at getting the level of robustness we need with such an approach. What happens is that this module is completely alert, is a neural network that is learned by trying it out in simulation and learning how to appropriately integrate these two modalities.

**VLADLEN:** And it's recurrent because it has a memory because it, it keeps an internal representation that persists through time. It is updated through time. So in the next timestamp, it doesn't wake up and see the world for the first time. It has an internal representation that it carries its mental state that it carries from the past. And it's just updating its mental states.

**CRAIG:** Yeah. And when I was asking before about the prior state-of-the-art, Boston Dynamics has done an excellent marketing job, so everyone's familiar with their quadrupedal robots. How do their controllers work? Are they combining proprioception and exteroception? Or are they primarily proprioception?

**VLADLEN:** Well, I have tremendous respect for Boston dynamics. Boston Dynamics is not just marketing, it's the real deal. Mark Raibert, has been setting the state of the art of four legged locomotion since the eighties. He's a true pioneer and visionary.

**VLADLEN:** And, we wouldn't be where we are without his commitment and his example, and Boston dynamics, many of the things they do, they're absolutely incredible. If you look at Spot, for example, the quadrupedal robot from Boston Dynamics, it does have visual perception and it does integrate visual perception and proprioception, but it does that using more classic methods.

**VLADLEN:** Where the integration framework hasn't been learned as much, but it relies much more on explicit, classic, more heuristic, robotics methodologies that were more popular in the past. So that predate, this learning-based methodology that we are advocating in our work. As a result, it seems to be considerably less robust and more brittle and more limited.

**VLADLEN:** For example, with Spot, it can go up and down stairs, but you have to manually switch it into stairs mode. You, the human operator, have to tell spot explicitly ‘Now Spot you are going to go up stairs.’ There's only one way you can take it upstairs. You have to position it straight where it's looking straight into the stairs and then has to go up straight in a straight line. Okay. No tilting. It's not going to do figure eights on the stairs. It's not going to rotate on this. It has to go up straight. Okay. It goes down straight, has to go down straight. Okay. If you look at our robot, first of all, you don't have to tell her anything.

**VLADLEN:** And these don't have to be well-defined stairs. It doesn't need stairs mode. If it finds itself being driven up stairs, it's going to go up stairs. It's going to go up stairs and in your orientation. It can dance on the stairs. It can do figure eights on the stairs. You don't need to tell it anything.

**VLADLEN:** It is basically… In our experiment, it never falls, no matter what the shape of the stairs is. And no matter what orientation and what trajectory and what pattern of motion it is taking. Much like a like a dog. Okay. It would take tremendous effort to make a dog trip and fall. Okay. With Spot, it does not take tremendous effort to make Spot trip and fall.

**VLADLEN:** You just, maybe don't tell it that is going up stairs and maybe oriented at a less than perfect orientation and drive it up stairs.

**CRAIG:** You use the word ‘drive’ and a lot of people don't realize that the, the, the robots that they see on YouTube, for example, are being controlled by a human controller or driven in effect by a human of screen. Two questions, is the model, your controller, on the edge devices, on the robot, does it require a link back to a central server or the cloud or something like that? And, how difficult is it to make those autonomous, or is that a completely different problem?

**VLADLEN:** There is no need for a cloud backend. The robot functions with the sensing and computation and actuation completely on board. It's basically a self-contained package as far as, as far as computation goes. Our robot, like Spot, needs to be navigated at a higher level generally by a human operator.

**VLADLEN:** Maybe a prerecorded set of instructions, but most commonly there is a human somewhere with a joystick, that says, well, go here, go there. The human is not controlling individual joints. It's not controlling the legs. And that really critical. The human is rather controlling the orientation and the trajectory of the robot at a high level,

**VLADLEN:** Making the robot autonomous so that you don't need the human around is an interesting problem. And, it is part of the program. It is somewhere that we want to get. We want to be able to, for example, tell the robot autonomously, go into this environment and map it out and maybe find something.

**VLADLEN:** So an incredible achievement would be if we have like a robot that can do what Alpine rescue dogs do. So there are dogs, actual physical, biological dogs that save lives by going out and looking for stranded hikers in the Alps. And they go out and autonomously explore the environment for hours and locate stranded hikers somewhere deep in the mountains and bring medical supplies and nourishment to them.

**VLADLEN:** These animals save lives. And it would be incredible if we can get this level of autonomy on this level of intelligence in our robots. I would love to get there. I'm personally striving to get there within my lifetime, within my career. I'd love to be part of that journey. And this higher level navigation, autonomous exploration aspect is not ready.

**CRAIG:** Can you talk a little bit about the flying of high-speed drones through complex environments? I can see that that relates very much to what you're doing with the quadrupedal robots, but it relies on a different, model. Can you tell us about that work?

**VLADLEN:** Yeah, that is a collaboration with a different lab in Zurich, University of Zurich, an incredible lab that likewise they do all the real work. And in that line of work, we are creating increasingly capable, increasingly autonomous, increasingly agile flying machines, flying machines that can at high speed traverse complex environments.

**VLADLEN:** And again, there, we need to perceive the environment because the drones need to avoid obstacles. So in some of our demonstrations, they fly through dense forest and avoid trees and they need to at very high speed, adapt to their actuation, the control signal, to what they perceive about the environment.

**VLADLEN:** So the overall setup of coupling perception and control in a closed loop, that is shared between the two lines of work. But the embodiments are very different.

**CRAIG:** With the drones, there's a higher level of autonomy. I mean, you're flying them toward a target, but they're making the decisions for their flight path through the obstacles.

**VLADLEN:** To some extent, yes, the drones are not driven by a human that kind of commands at a high level where the drone needs to go. In a typical deployment that drones are given a rough reference trajectory through the environment, but that trajectory may not be adapted to the environment and it may not avoid obstacles.

**VLADLEN:** So if the drone just executes the trajectory as it was given, it's just going to die. It's just going to crash, crash into a tree. So a typical task would be, follow the trajectory to the best extent you can get to the goal or as close to the final point on the trajectory as you can, but also survive. Don't hit anything. don't collide with any obstacles.

**CRAIG:** The other work that you've done that I'm really fascinated by is this high speed, computer vision training. I think it's limited to computer vision, where you have, is it a million frames per second? Some incredible number. I think the platform that you built that does this is called the Megaverse, a 3D simulation platform. Was that used in the training of the drones or the training of the quadrupedal robots? And if not, what applications have you trained on the Megaverse?

**VLADLEN:** The Megaverse platform is a platform for high throughput reinforcement learning, high throughput training of embodied intelligence in simulation with high fidelity visual perception. So there, we simulate agents that interact with physical environments and the agents act based on high fidelity images, the kinds of images our retinas or cameras form.

**VLADLEN:** So, we simulate light transfer the way light interacts with surfaces and we produce color images. The kinds of images you and I are familiar with. And this is done in the Megaverse at on the order of a million frames per second. And why do we care about being so fast, simulating so much faster than real time.

**VLADLEN:** The basic observation is that if you look at humans, for example, it takes human some years to get to a level of physical embodied intelligence, a level of competence that we would like to see in intelligent systems. Depending what scale your track and what physical competence you care about, it takes at least a year, let's say. Safe to say at least a year of embodied experience in the world to tune all the motor controllers and the coupling of perception and action to tune the perception system, to calibrate the motor control, and to couple. So let's say at least a year. And the human is a much more effective learning system than anything we have in the lab, than any artificial learner that we have designed so far, by, depending how you count, some orders of magnitude.

**VLADLEN:** Let's say one or two orders of magnitude. Our learning frameworks are very simple, intensive. They need a tremendous amount of experience. Then you need a tremendous amount of data. So let's say some orders of magnitude more. So that tells you that maybe you need at least on the order of 10 to a hundred years of experience in simulation to learn the kinds of motor skills that, maybe a one year old or a two year old human child have. But that's not all because we also need to design the learner. We need to do this many, many, many times to try out different learning architectures, different neural network architectures, different learning paradigms, learning protocols, curricula, different environments, different sequences of tasks that we take the learner through.

**VLADLEN:** So there is this outer loop. That means that you need to go through this 10 to a hundred year learning process many, many, many, many times. Now we want to do this, such that this iteration process can be done in hours or days, not years. I don't want to launch an experiment. And then we say, well, Craig, let's talk in 2032 and I'll tell you how the experiment went.

**VLADLEN:** I need to be able to do this basically every day in the lab and see experimental results within hours. That's why we build platforms like Megaverse that allow us to simulate years of subjective experience in hours.

**CRAIG:** And have you used that? What models have you tried? Did you use that in either of the two papers that we just talked about?

**VLADLEN:** Megaverse actually came after. It's a recent project. And it is informed by some work we did in the other lines of work. But for example, for legged locomotion, we used an in-house boutique high-performance simulation framework that was built at ETH Zurich. That was highly optimized specifically for these legged systems.

**VLADLEN:** And that was one of the inspirations in that framework. We didn't really need high fidelity images with pixels. We didn't need color images because the legged systems didn't look at color images in order to locomote. With Megaverse, we are pushing into more advanced, higher level cognitive tasks where the agent needs to interact with the environment, manipulate it, maybe build a bridge over a gap or dismantle some kind of obstacle or move some objects into a specific configuration or assemble a tower, advanced, embodied tasks that involve manipulation and the interaction.

**VLADLEN:** And for which visual perception, high-fidelity visual perception are important. So one of the big accomplishments in Megaverse is to be able to simulate high fidelity visual perception with full color images at these incredible framework.

**CRAIG:** And that was specifically for reinforcement learning, is that right?

**VLADLEN:** Yeah. So, that was the learning approach we had in mind when we designed and built the platform.

**CRAIG:** But it could also be used for supervised learning, couldn’t it, and with virtual data, I mean, virtually generated images.

**VLADLEN:** Yes, you could definitely use it for supervised learning. It will allow you to collect data at incredible frame rates and learn very fast. Yes, it absolutely can be used in a supervised learning setup.

**CRAIG:** And, and so where is that, where's Megaverse going? Is this a field, high throughput frame rate, that, many people are working on? Is this unique? And where is it going. Is it going to be applied more generally? It just seems like, a very natural, idea for training models.

**VLADLEN:** Yes, I think this way of thinking and some of the techniques that we introduced are being adopted and will be increasingly deployed by other teams as well.

**CRAIG:** That’s it for this week’s episode. Again, I want to thank [ClearML](https://t.clear.ml/eye_on_ai) for their support. Please take a moment and visit [clear.ml](https://t.clear.ml/eye_on_ai) to see what they offer. At the very least they’ll see that sponsoring the podcast is worth their while.

I want to thank Vladlen for his time. We’ll be how the Megaverse expands and how robotic intelligence unfolds.

In the meantime, remember, AI is about to change your world, so pay attention.

**VLADLEN:** I'm aware of some projects that are not public yet that are building on our work. I think the goal of simulating at pretty high fidelity, at orders of magnitude faster than reality, is a goal that is shared by many groups and many groups are working in that direction.

**CRAIG:** Is Megaverse available, is it open source, available for people to use? I mean, do you have to apply for time on the Megaverse or how does that work?

**VLADLEN:** Oh, it's, it's available free, open source with a permissive license, very easy to use. And if you encounter any issues, just submit a GitHub issue and we'll look at it. Yeah. We definitely are very happy when people use it.

**CRAIG:** So it's a platform. It is one platform, or it is an architecture that people can take and build in their own labs.

**VLADLEN:** Both. So people can compile the code and run Megaverse on their servers, this Megaverse that we built with the tasks that we provided. So people can reproduce our work and kind of incrementally build on it.

**VLADLEN:** That's one. And other scenarios that people can read our code and see how we did what we did. And look at the ideas described in the paper. And there are a number of systems, architecture patterns, design patterns that enabled us to get to these incredible levels of performance and people can adopt them in building a different platform that maybe has other kinds of agents, other kinds of environments, other kinds of tasks.

**CRAIG:** And, where is all this going in your mind? You've got the sort of tight loop between physical locomotion and perception. You've got this super high speed, training platform. Where are you pushing in your research?

**VLADLEN:** Well, I think a major evolution is going to be when we have physical systems that effectively interact with environments and manipulate environments to achieve certain goals.

**VLADLEN:** If you look so far at our work with legged systems and with drones, the accomplishments so far are in having these systems effectively traverse environments in an agile sometimes graceful way. They don't fall. They avoid obstacles, they don't crash. They get to where we need them to be. But of course you and I do much more with physical environments.

**VLADLEN:** We manipulate objects, we rearrange the environment to fit our goals. And that is a level of engagement that we haven't tackled so far in most of our work. And I think that's going to be the next era, the next phase in embodied AI.