Received January 24, 2022, accepted March 4, 2022, date of publication March 10, 2022, date of current version March 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158758

On the Scalability of Vision-Based Drone Swarms in the Presence of Occlusions

FABIAN SCHILLING, ENRICA SORIA, AND DARIO FLOREANO, (Senior Member, IEEE)
Laboratory of Intelligent Systems, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Corresponding author: Fabian Schilling (fabian.schilling@epfl.ch)

This work was supported by the Swiss National Science Foundation under Grant 200021-155907.

ABSTRACT Vision-based drone swarms have recently emerged as a promising alternative to address the fault-tolerance and flexibility limitations of centralized and communication-based aerial collective systems. Although most vision-based control algorithms rely on the detection of neighbors, they usually neglect critical perceptual factors such as visual occlusions and their effect on the scalability of the swarm. To estimate the impact of occlusions on the detection of neighbors, we propose a simple but perceptually realistic visual neighbor selection model that discards obstructed agents. We evaluate the visibility model using a potential-field-based flocking algorithm with up to one thousand agents, showing that occlusions have adverse effects on the inter-agent distances and velocity alignment as the swarm scales up, both in terms of group size and density. In particular, we find that small agent displacements have considerable effects on neighbor visibility and lead to control discontinuities. We show that the destabilizing effects of visibility switches, i.e., agents continuously becoming visible or invisible, can be mitigated if agents select their neighbors from adjacent Voronoi regions. We validate the resulting flocking algorithm using up to one hundred agents with quadcopter dynamics and subject to sensor noise in a high-fidelity physics simulator. The results show that Voronoi-based interactions enable vision-based swarms to remain collision-free, ordered, and cohesive in the presence of occlusions. These results are consistent across group sizes, agent number densities, and relative localization noise. The source code and experimental data are available at https://github.com/lis-epfl/vmodel.

INDEX TERMS Unmanned aerial vehicles, multi-robot systems, agent-based modeling, scalability, vision.

I. INTRODUCTION

Aerial robot swarms have a vast socio-economic potential and are used for numerous real-world applications in industries such as agriculture, mapping, and construction [1]–[3]. Drone swarms can be deployed to monitor crops, create maps, and survey sites much faster than a single drone since they can solve tasks cooperatively and in parallel. Larger group sizes can further decrease task completion times and operating swarms in compact formations can enable new applications in confined spaces such as buildings. However, most drone swarms deployed today rely on external localization and wireless communication, both of which represent major limiting factors towards their scalability in terms of group size and swarm density.

Localization in drone swarms is usually achieved with satellite-based systems for outdoor applications [4] or optical motion capture for indoor deployments [5]. The drones are typically equipped with wireless communication devices that enable the exchange of state information such as positions and velocities with each other. While this approach has enabled successful deployments of impressive aerial swarms, it comes with several key limitations. Firstly, wireless communication suffers from inherent scalability issues since the bandwidth requirement scales quadratically with the number of agents [6]. In practice, this leads to compounding delays whose durations are difficult to estimate and thus require dampening and interpolation [7], [8]. Secondly, the approach lacks flexibility since the agents must adhere to the same communication protocol and need to be localized in the same frame of reference. Thirdly, the exclusive use of an external positioning system represents a single point of failure and its malfunctioning can have disastrous effects.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

Vision-based relative localization methods rely entirely on local information to detect other agents, thus removing the dependence on external localization systems and additional communication devices (e.g., Bluetooth [9], ultra-wideband [10], and others [11]). Moreover, vision is arguably the ideal sensory modality for localization on aerial robots since cameras are small, lightweight, and provide extremely high information density at comparatively low power consumption [12]. Multi-robot systems that use a vision-based approach to mutual localization have recently emerged in the form of leader-follower formations [13]–[15] and the first aerial flocks [16]–[18]. Important perceptual factors such as visual occlusions, i.e., agents that are obstructed by others, are usually neglected in these swarms because of their small group size. However, these factors become a deterrent for larger swarms, especially when they have to fly in dense configurations.

While some swarm roboticists explicitly make use of visual occlusions to solve collaborative transport problems [19] and robotic shepherding tasks [20], the most thorough treatment of visibility constraints can be found in the collective motion literature. Using computer vision techniques, researchers are able to reconstruct the poses and visual fields of individual animals and show that visual perception best explains how information about food sources and predators transfers within the group [21]–[24]. How individuals select and react to their neighbors is one of the fundamental questions in the study of collective motion and agent-based flocking models provide an indispensable tool to test and verify different hypotheses [25]–[28]. Notable examples of neighbor selection methods include *metric* (i.e., within a metric radius) [29], *topological* (i.e., the set of n nearest neighbors) [30], or *voronoi*-based (i.e., from adjacent Voronoi regions) [31] interactions. Recently, different forms of *visual* neighbor selection have gained popularity due to their biological plausiblity [21]–[24]. For example, research on flocking models with a limited field of view shows that lateral vision is crucial for collision-free collective motion [32], [33] and may explain why flocking birds have almost omnidirectional vision [34]. Simulations of large schools of fish show that visual obstructions lead to more realistic group shapes and densities than purely metric interactions [35]. Simulations of large vision-based flocks show that bird density can be regulated effectively if individuals only react to the projection of their neighbors [36]. Other researchers show that many natural behaviors such as milling and polarized flocking emerge from purely visual interactions even in the absence of a spatial representation of neighbors [37]. Although these models offer interesting collective behaviors, they often make modeling choices that are geared towards a particular species or result in undesirable behavior for robotic swarms since they lead to frequent collisions.

In this work, we tackle visibility constraints arising from occlusions from a robotics perspective with the goal of synthesizing large and compact vision-based drone swarms. In particular, we study the effect of occlusions on the performance (i.e., collision avoidance, cohesion, and velocity alignment) of vision-based swarms as they scale from low densities and a handful of agents to high-density swarms with thousands of individuals. To this end, we propose a *visual* neighbor selection model that offers a perceptually plausible alternative to the ubiquitous but unrealistic *metric* selection of neighbors, i.e., methods that assume agents can sense arbitrary neighbors within a given radius regardless of occlusions. We simulate vision-based swarms of up to one thousand point mass agents and program them to perform collective waypoint navigation using a simple attractive/repulsive flocking algorithm. The results show that swarms in which agents react to all *visible* neighbors perform poorly, especially at high densities and as the group size increases beyond tens of agents. However, by limiting visual interactions to their Voronoi neighbors, we can successfully synthesize collision-free, cohesive, and ordered vision-based swarms. A comparison of Voronoi interactions with other common neighbor selection methods (i.e., metric and topological) reveals their superiority in large, high-density swarms. We validate the scalability of the resulting flocking algorithm at different densities and group sizes with quadcopter dynamics using a simulator with realistic physics and noise levels. The analysis shows that visually-constrained Voronoi interactions are both perceptually plausible and highly effective for the coordination of large aerial robot swarms in which agents rely purely on local visual information for control.

II. METHOD

We aim to synthesize a vision-based swarm that remains as compact as possible and collision-free while performing collective waypoint navigation. We define this objective since it enables many practical applications such as cooperative mapping, aerial deliveries, and search & rescue.

We briefly define preliminary concepts and the notation used throughout the article (Sec. II-A). We then describe a simple attractive/repulsive flocking algorithm that provides collision avoidance and cohesion, as well as a navigation capability to the swarm (Sec. II-B). To obtain a flocking algorithm that is plausible for vision-based swarms, we define the notion of agent visibility in the form of a neighbor selection strategy that is based on a realistic occlusion model (Sec. II-C). Since vision-based detection is an inherently stochastic process, we further model sensing noise on the range and bearing measurements (Sec. II-D).

A. PRELIMINARIES AND NOTATION

We consider a set of N homogeneous agents that are labeled by $i \in \mathcal{A}$, where $\mathcal{A} = \{1, 2, \ldots, N\}$ denotes the set of all agents and $|\mathcal{A}| = N$ its cardinality. We also define the set of all but the focal agent i as $\mathcal{A}_i = \mathcal{A} \backslash \{i\}$. The state of each agent i can be described by its position and velocity $\mathbf{p}_i, \mathbf{v}_i \in \mathbb{R}^m$. We focus on the two-dimensional case and let $m = 2$, assuming that the agents move in planar configurations. We denote the relative position of agent i with respect to j as $\mathbf{r}_{ij} = \mathbf{p}_j - \mathbf{p}_i$ with distance $d_{ij} = \left\|\mathbf{r}_{ij}\right\|$ where $\|\cdot\|$ is the Euclidean norm.

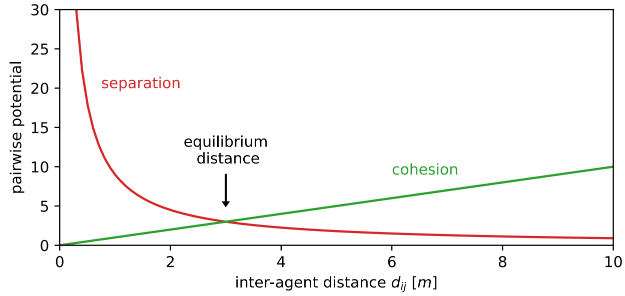


FIGURE 1. Pairwise potential of separation and cohesion terms as a function of inter-agent distance. Separation is inversely proportional to the inter-agent distance, whereas the cohesion term grows linearly with distance. The equilibrium distance is defined as the distance at which separation and cohesion balance.

We model the swarm of agents as a directed sensing graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices $\mathcal{V} = \{1, \ldots, N\}$ denotes the agents and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains the ordered pairs of agents $(i, j) \in \mathcal{E}$ if an agent i is adjacent to agent j, which we denote by $i \sim j$. The graph $\mathcal{G}$ can also be represented by an $N \times N$ adjacency matrix of the form A_{ij} with entries of 1 if $i \sim j$ and 0 otherwise.

The motion of each agent can be described by single-integrator dynamics of the form

$$\mathbf{p}_i^{k+1} = \mathbf{p}_i^k + \mathbf{v}_i^k \Delta t \tag{1}$$

where k denotes the index of the discrete time step with duration Δt.

In the remainder of the section, we skip the dependence on the discrete time step k for notational brevity and clarity. However, all computations in this section are performed at every time step without exception.

B. FLOCKING ALGORITHM

The objective of the swarm is to perform waypoint navigation while avoiding inter-agent collisions and staying together as a group. We formulate this objective as an artificial potential field that is inspired by the Reynolds flocking algorithm [38]. The motion of an agent is composed of an attractive/repulsive potential that provides separation and cohesion between agents (Sec. II-B1), as well as a migratory potential responsible for goal-directed navigation (Sec. II-B2).

The motion of an agent is composed of a social term that captures agent-to-agent interactions and a migration term that introduces the navigation objective. The velocity command of an agent can be written as

$$\mathbf{v}_i = \mathbf{v}_i^{\text{soc}} + \mathbf{v}_i^{\text{mig}} \tag{2}$$

where $\mathbf{v}_i^{\text{soc}}$ and $\mathbf{v}_i^{\text{mig}}$ denote the respective social (Eq. 3) and migration terms (Eq. 4). In order to obtain a final velocity command that is feasible even under the actuation constraints of a physical robot, we limit the maximum speed as $\tilde{\mathbf{v}}_i = \mathbf{v}_i / \|\mathbf{v}_i\| \min(\|\mathbf{v}_i\|, v^{\max})$. The velocity command $\tilde{\mathbf{v}}_i$ can then be used directly for the motion update to obtain the agent positions of the next time step (Eq. 1).

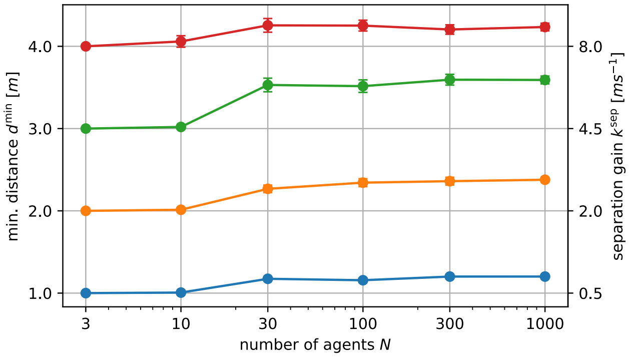


FIGURE 2. Scalability of minimum nearest neighbor distances to increasing numbers of agents using the baseline *metric* neighbor selection model, i.e. agents within the perception radius are detected irrespective of whether they are occluded. Each line represents the minimum equilibrium distance between nearest neighbors obtained from different separation gains as the swarm size increases (mean and std. dev. over ten trials, all other parameters constant). Aside from a noticeable increase of inter-agent distances between ten and thirty agents that occurs due to the saturation of the perception range with agents, the inter-agent distances remain constant across different group sizes (note the logarithmic scale).

1) SEPARATION AND COHESION

Cohesion and collision avoidance can be achieved with an attractive/repulsive potential that keeps the agents at an equilibrium distance (Fig. 1). The cohesion term keeps the swarm together by attracting agents to the average position of their neighbors. The separation term leads to collision avoidance by repulsing nearby agents from each other. We can express these rules more formally as

$$\mathbf{v}_i^{\text{soc}} = k^{\text{coh}} \underbrace{\frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{r}_{ij}}_{\text{cohesion}} - k^{\text{sep}} \underbrace{\sum_{j \in \mathcal{N}_i} \frac{\mathbf{r}_{ij}}{\|\mathbf{r}_{ij}\|^2}}_{\text{separation}} \tag{3}$$

where k^{coh} and k^{sep} are gains that regulate the strength of the attraction and repulsion, respectively.

Note that we do not scale the separation velocity command by the number of agents. This formulation has the advantage that minimum inter-agent distances remain quasi-constant as the group size increases and thus reduces the need for readjusting the control gains (Fig. 2). We further use the analytical solution to the above equations for three agents as a first approximation of the desired inter-agent distance d^{ref}. This allows us to express an approximate reference distance by using a separation gain of the form $k^{\text{sep}} = (d^{\text{ref}})^2/2\,\text{m}\,\text{s}^{-1}$ and keeping the cohesion gain fixed at $k^{\text{coh}} = 1\,\text{m}\,\text{s}^{-1}$. Note that in general, the separation gain slightly overestimates the reference distance for larger swarms since it does not take the number of neighbors into account. It is nevertheless a useful approximation that spares us the tedious task of finding the reference distance empirically for each agent swarm scale separately.

2) MIGRATION

The purpose of the migration term is to give the agents a navigation goal by steering them towards a waypoint. The

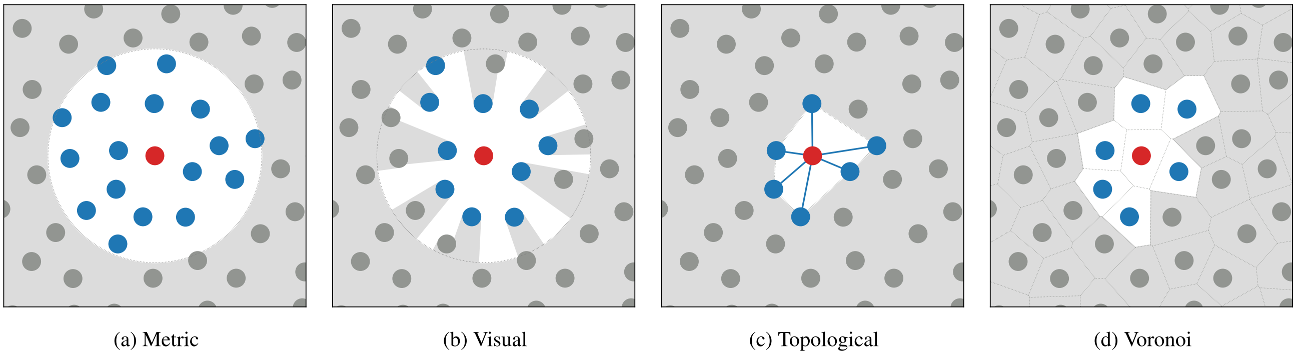

(a) Metric (b) Visual (c) Topological (d) Voronoi

FIGURE 3. Schematic visualization of different neighbor selection strategies: metric, visual, topological, and voronoi. We take the perspective of a focal agent within a swarm (central red disk) that selects agents (blue disks) and discard others (gray disks) depending on the following selection criteria: (a) *metric* selects all agents within a metric perception radius, (b) *visual* selects all visible agents within a metric radius, i.e., all agents that appear large enough and are not occluded by others, assuming agents are equally sized and have an omnidirectional camera at their center, (c) *topological* selects only the *n* closest agents (here *n* = 6), irrespective of their distance, and (d) *voronoi* selects only those agents that belong to a neighboring Voronoi region.

migration term can be written as

$$\mathbf{v}_i^{\text{mig}} = k^{\text{mig}} \frac{\mathbf{r}^{\text{mig}}}{\left\| \mathbf{r}^{\text{mig}} \right\|} \tag{4}$$

where $\mathbf{r}^{\text{mig}}$ denotes the relative position of the migration point with respect to the focal agent, and k^{mig} the gain for modulating the migration speed.

C. NEIGHBOR SELECTION

Neighbor selection is an important consideration for all flocking algorithms since it introduces the notion of locality (e.g., in communication, perception, etc.) as opposed to all-to-all information transfer. In the following, we denote the neighbors of agent i as a set $\mathcal{N}_i$ where $\mathcal{N}_i \subseteq \mathcal{A}_i$.

1) METRIC: DISTANCE-BASED NEIGHBOR SELECTION

Metric neighbor selection keeps only those agents that fall within a radius $r^{\max}$ centered around the focal agent (Fig. 3a). We can formalize metric neighbor selection as the set

$$\mathcal{N}_i^{\text{metric}} = \left\{ j \in \mathcal{A}_i \mid d_{ij} < r^{\max} \right\} \tag{5}$$

where $r^{\max}$ denotes the maximum perception range.

Defining the set of neighbors based on a metric range is the most popular means of neighbor selection in the literature [29], [38]–[40]. Metric neighbor selection is a simple and effective method to introduce locality in the interactions and can be interpreted as a perception radius for vision-based swarms or a communication range for swarms that can exchange information via wireless links, for example. With the assumption that all agents are homogeneous and equally sized, we can use the metric perception range to represent visual acuity, i.e., the minimum size that another agent spans on the retina of the focal agent before it can no longer be perceived. We therefore encourage the reader to think about the perception range as the equivalent of the minimum subtended angle that another agent spans on the retina of the focal agent.

2) VISUAL: OCCLUSION-BASED NEIGHBOR SELECTION

Visual neighbor selection keeps only those agents that appear large enough and are not occluded by closer ones as seen from the perspective of the focal agent (Fig. 3b). The set of visible agents can be formalized as

$$\begin{aligned} \mathcal{N}_i^{\text{visual}} = \{ & j \neq k \in \mathcal{N}_i^{\text{metric}} \mid \\ & \neg \left(\left\| \mathbf{u}_{ij} - \mathbf{u}_{ik} \right\| < \hat{r}_{ij} + \hat{r}_{ik} \wedge d_{ij} < d_{ik} \right) \} \end{aligned} \tag{6}$$

where $\mathbf{u}_{ij} = \mathbf{r}_{ij}/d_{ij}$ and $\hat{r}_{ij} = r/d_{ij}$ are the projections of the agent position and radius onto the unit circle, respectively.

Note that by combining metric and visual neighbor selection, we obtain a model of visibility that takes into account both visual acuity and occlusions. We consider this model plausible for vision-based swarms since it captures the information that is de facto available to an individual that operates purely on visual perception.

The above definition of visibility contains two key assumptions. The first assumption is that agents can distinguish individuals from each other. Note that this assumption does not require identities to be maintained over time. The second assumption is that partially occluded agents are considered invisible, i.e., only the closest set of agents with an uninterrupted line of sight are contained in the visible set. This assumption is reasonable for monocular vision since the relative distance to other agents can only be reliably estimated if their entire spatial extent is visible.

3) TOPOLOGICAL: *n*-NEAREST NEIGHBOR SELECTION

Topological neighbor selection keeps only the n nearest neighbors of the focal agent (Fig. 3c). We can write the set of nearest neighbors as

$$\mathcal{N}_i^{\text{topo}} = \left\{ n - \underset{j \in \mathcal{A}_i}{\arg\min}\, d_{ij} \right\} \tag{7}$$

where the $n - \arg\min$ operator selects at most the n nearest neighbors.

Topological neighbor selection is a popular method due to its explanatory success in natural swarms [21], [41] and is often used in models of collective motion to maintain group cohesion [30], [40].

4) VORONOI: SPATIALLY BALANCED NEAREST NEIGHBOR SELECTION

Voronoi neighbor selection keeps only those agents whose Voronoi regions share a border with the focal agent (Fig. 3d). We can write the set of Voronoi neighbors as

$$\mathcal{N}_i^{\text{voronoi}} = \left\{ j \in \mathcal{A}_i \mid V_i \cap V_j \neq \emptyset \right\} \tag{8}$$

where ∅ denotes the empty set and V_i the Voronoi region of agent i which can be defined as

$$V_i = \left\{ j \in \mathcal{A}_i, \mathbf{q} \in \mathbb{R}^m \mid \|\mathbf{q} - \mathbf{p}_i\| \leq \left\| \mathbf{q} - \mathbf{p}_j \right\| \right\}. \tag{9}$$

In other words, the Voronoi region of an agent can be described as the set of all points that are closer to itself than to any other agent.

Neighbor selection based on the Voronoi tessellation can be seen as topological interactions that are parameter-free and automatically balanced in space [30]. Moreover, it can be shown that the average number of Voronoi neighbors is at most six for the planar case we are considering here [42].

D. SENSING NOISE

We model the visual relative localization inaccuracies in two independent components: range and bearing. We model range noise as a function that varies linearly with relative distance from the observer whereas the bearing noise is constant over the field of view [15], [18], [43], [44]. More formally, we define the noisy version of range and bearing with which agent i detects agent j as

$$\hat{d}_{ij} = d_{ij}(1 + \omega_d), \quad \omega_d \sim \mathcal{N}(0, \sigma_d) \tag{10}$$

$$\hat{\beta}_{ij} = \beta_{ij} + \omega_\beta, \quad \omega_\beta \sim \mathcal{N}(0, \sigma_\beta) \tag{11}$$

where ω_d and ω_β are independent and identically distributed white noise with zero mean and standard deviation of σ_d and σ_β, respectively. The noisy relative position can then be constructed from polar coordinates as

$$\hat{\mathbf{r}}_{ij} = \begin{bmatrix} \hat{d}_{ij} \cos(\hat{\beta}_{ij}) \\ \hat{d}_{ij} \sin(\hat{\beta}_{ij}) \end{bmatrix} \tag{12}$$

where $\hat{\mathbf{r}}_{ij}$ can serve directly as an input to the social term of the flocking algorithm (Eq. 3). The exact values for range and bearing noise depend on several factors such as camera resolution, lens quality, calibration accuracy, and target deformation.

III. EXPERIMENTAL SETUP

Before we analyze the experimental results, we briefly describe the metrics that we use to measure the swarm performance (Sec. III-A), as well as the experimental parameters that are used throughout the experiments (Sec. III-B).

A. PERFORMANCE METRICS

We report our results in terms of several complementary metrics: minimum nearest neighbor distance $d^{\min}$, order ϕ^{order}, and union ϕ^{union}. These metrics capture whether we have achieved collision-free, aligned, and cohesive collective navigation, respectively. The following metrics are computed at every discrete time step k and we therefore omit the time-dependence for notational brevity.

The minimum nearest neighbor distance is arguably the most important metric since it captures whether or not the agents can effectively avoid collisions during migration. It is computed as

$$d^{\min} = \min_{i \neq j} d_{ij} \tag{13}$$

and we say that a collision occurs whenever two agents get closer than twice their radius $d^{\min} < 2\,r$.

The order metric measures the correlation of the velocity vectors of the agents within the swarm. It is computed as

$$\phi^{\text{order}} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \left\|\mathbf{v}_j\right\|}. \tag{14}$$

An order value of one indicates that all agents are moving in the same direction in perfect alignment, whereas a value around zero means that the swarm is in a completely disordered state in which no two agents align their direction of motion.

The union metric measures the cohesion of the swarm and expresses whether the swarm has split into subgroups. It is computed as

$$\phi^{\text{union}} = 1 - \frac{n^{\text{comp}} - 1}{N - 1} \tag{15}$$

where n^{comp} is the number of connected components of the neighbor adjacency matrix (Sec. II-C). A union value of one indicates that the swarm is moving as a single cohesive unit. A value of zero represents the degenerate situation in which the swarm is split into N subgroups and the agents are unable to perceive any other agent.

B. EXPERIMENTAL PARAMETERS

We perform ten repeated runs of migration experiments to make statistical statements about the scalability of the swarm using different neighbor selection methods, group sizes, swarm densities, agent dynamics, and noise levels.

The specific parameter values we use are informed by our previous experiments with real vision-based quadcopters in indoor [14] and outdoor environments [18], as well as the literature on vision-based drone localization [13], [16], [43]–[49]. We choose the radius of an agent as $r = 0.25$ m since it reflects a common physical size of quadcopter platforms used in robotic experiments. The perception radius $r^{\max} = 10$ m is chosen as the distance at which other drones were no longer reliably detected during outdoor experiments. The time delta $\Delta t = 100$ ms is chosen as a reasonable amount of time to solve the visual perception, state estimation, and control

TABLE 1. Neighbor selection methods used during the experiments.

Name	Set notation
metric	$\mathcal{N}_i^{\text{metric}}(r^{\max})$
visual	$\mathcal{N}_i^{\text{visual}}(r^{\max})$
visual + myopic	$\mathcal{N}_i^{\text{visual}}(2d^{\text{ref}})$
visual + topological	$\mathcal{N}_i^{\text{visual}}(r^{\max}) \cap \mathcal{N}_i^{\text{topo}}(n)$
visual + voronoi	$\mathcal{N}_i^{\text{visual}}(r^{\max}) \cap \mathcal{N}_i^{\text{voronoi}}$

TABLE 2. Parameters used during the experiments.

Description	Notation	Value
Agent radius	r	$0.25\,\text{m}$
Reference distance	d^{ref}	$1\,\text{m}$
Perception radius	$r^{\max}$	$10\,\text{m}$
Bearing noise	σ_β	1°
Range noise	σ_d	$0.05\,\text{m}$
Maximum topological neighbors	n	6
Maximum speed	$v^{\max}$	$1\,\text{m}\,\text{s}^{-1}$
Separation gain	k^{sep}	$1\,\text{m}\,\text{s}^{-1}$
Cohesion gain	k^{coh}	$1\,\text{m}\,\text{s}^{-1}$
Migration gain	k^{mig}	$0.5\,\text{m}\,\text{s}^{-1}$
Time delta	Δt	$0.1\,\text{s}$
Simulation duration	T	$200\,\text{s}$

problems in real-time. The desired inter-agent distance is set to $d^{\text{ref}} = 1\,\text{m}$ to generate the most compact formation that simultaneously provides enough safety margin against potential collisions.

In order to provide a fair comparison of the visual neighbor selection methods, we choose parameter values that result in comparable numbers of neighbors as the group size increases (Fig. 4d). In particular, we set the maximum number of agents for topological neighbor selection to $n = 6$ since it reflects the average number of Voronoi neighbors for planar configurations [42]. We further let $r^{\max} = 2d^{\text{ref}}$ for myopic interactions since it approaches an average number of six neighbors as the group size increases. We provide an overview of the neighbor selection methods used during the experiments in Tab. 1.

Note that *metric* neighbor selection is not plausible for swarms in which relative localization is vision-based. We include an analysis of *metric* neighbor selection only for comparison and because it is commonly used in the literature. Conversely, all other *visual* neighbor selection methods (i.e., *visual*, *visual* + *myopic*, *visual* + *topological*, *visual* + *voronoi*) are feasible for vision-based swarms since the agents have uninterrupted line of sight.

At the beginning of each experiment, the agents are spawned randomly within a circular region. The initial positions are sampled uniformly in a non-overlapping fashion using rejection sampling such that no pair of agents are closer than their desired reference distance d^{ref}. The area of the circular region is chosen such that the agent number density ρ_N remains constant for different numbers of agents. The agents exhibit no motion at the beginning of the experiment, i.e., their initial velocities are set to zero. The agents are given a constant navigation direction $\mathbf{r}^{\text{mig}} = [1, 0]^\top$ along the horizontal axis which can be seen as a migratory route along the magnetic field [50]. We let the swarm develop its collective motion for a total of $T = 200\,\text{s}$ composed of 2000 isochronous discrete time steps k with duration $\Delta t_k = 0.1\,\text{s}$. At each time step, the agents select their neighbors according to the indicated neighbor selection function (Fig. 3) and compute their motion command (Sec. II-B). We set the separation and cohesion gains to $k^{\text{sep}} = 1\,\text{m}\,\text{s}^{-1}$ and $k^{\text{coh}} = 1\,\text{m}\,\text{s}^{-1}$ to provide an approximate nearest neighbor distance of $d^{\text{ref}} = 1\,\text{m}$. The separation gain is set to $k^{\text{mig}} = 0.5\,\text{m}\,\text{s}^{-1}$ which provides goal-directed motion without overpowering the attractive/repulsive commands. We set the maximum speed an agent can sustain to $v^{\max} = 1\,\text{m}\,\text{s}^{-1}$. A concise overview of the experimental parameters is provided in Tab. 2.

In order to provide a fair comparison across vastly different group sizes, we compute the metrics over the last quarter of the simulation, i.e. considering only the final 500 time steps. Particularly for large swarm sizes, we avoid computing metrics during an initial transient period in which agents have not yet aggregated to their final configuration. We refer to the time range during which we compute the metrics as the equilibrium period for convenience.

We report the minimum nearest neighbor distances as a minimum over time over the equilibrium period since it reveals whether collisions occur. For the order and union metrics, we report time averages over the equilibrium period. The mean and standard deviations are computed over the ten independent runs with random initial conditions.

C. SIMULATION ENVIRONMENTS

We employ two different simulation environments that serve complementary purposes. The simulation environment with point mass dynamics allows us to rapidly prototype algorithms and quickly generate statistical results with up to one thousand agents without running into time or computational constraints.

The Gazebo simulator, on the other hand, provides more physical realism and allows us to obtain an approximation of how an algorithm would behave on real hardware. However, by default, Gazebo, ROS, and PX4 run asynchronously, meaning that messages are exchanged on a best-effort basis given the computational load. To provide a fair comparison at different group sizes, we must ensure that the number of agents does not have any adverse effects on the simulation fidelity by lockstepping all of its software components. In practice, this means we run Gazebo and PX4 in their respective lockstep modes and additionally pause the simulation at each time step, compute the velocity commands for all agents in parallel, and resume the simulation. Unfortunately, even with lockstepping, Gazebo reaches its computational limits at around one hundred agents, after which the real-time factor decreases considerably and spawning additional agents becomes unreliable. We therefore limit our experiments with quadcopter dynamics to one hundred agents.

IV. RESULTS

We report results on four sets of complementary simulation experiments: 1) we compare several neighbor selection methods with increasing numbers of agents to show their performance for different swarm sizes (Sec. IV-A), 2) we evaluate the neighbor selection methods for increasing inter-agent distances to show the effect of varying agent number densities on the swarm performance (Sec. IV-B), 3) we analyze the performance of the neighbor selection methods when they are subjected to increased range noise during relative localization (Sec. IV-C), and 4) we validate the highest-performing neighbor selection method (across group sizes and densities) with quadcopter dynamics and realistic sensing noise to show its performance under real-world conditions (Sec. IV-D).

A. PERFORMANCE ACROSS SWARM SIZES

We assess the performance of the swarm for all neighbor selection methods and six levels of increasing group size $N \in \{3, 10, 30, 100, 300, 1000\}$. We set the reference distance $d^{\text{ref}} = 1\,\text{m}$ constant throughout the experiments to keep the agent number density fixed and to allow a direct comparison of the effect of group size.

1) VISUAL NEIGHBOR SELECTION

Purely *visual* neighbor selection shows the overall lowest performance as the group size increases. There is a considerable performance penalty in the distance and order metrics (Fig. 4a and 4b). The minimum distance is tracked well only for a group size of 3 agents ($d^{\text{min}} = 1.0 \pm 0.0\,\text{m}$; Fig. 4a). The distance gradually approaches the collision threshold of $2r = 0.5\,\text{m}$ and reaches its minimum at 1000 agents ($d^{\text{min}} = 0.58 \pm 0.0\,\text{m}$; Fig. 4a). The order metric shows a similar trend since the agents start out perfectly ordered for 3 agents ($\phi^{\text{order}} = 1.0 \pm 0.0$; Fig. 4b). However, for larger group sizes, the order metric decreases monotonously until reaching its minimum at 1000 agents ($\phi^{\text{order}} = 0.87 \pm 0.0$; Fig. 4b). The swarm stays cohesive as a single unit across all group sizes ($\phi^{\text{union}} = 1.0 \pm 0.0\,\text{m}$; Fig. 4c). Generally, using *visual* neighbor selection, the swarm performance decreases as soon as occlusions start to emerge (Fig. 4d; Fig. 4d). There is no performance penalty for 3 agents using *visual* neighbor selection since they predominantly occur in equilateral triangle formations in which there are no occlusions (i.e., $N_i = 2$). For larger group sizes, an increasing number of agents within the perception radius is occluded (32% occluded for $N = 10$; up to 90% occluded for $N = 1000$).

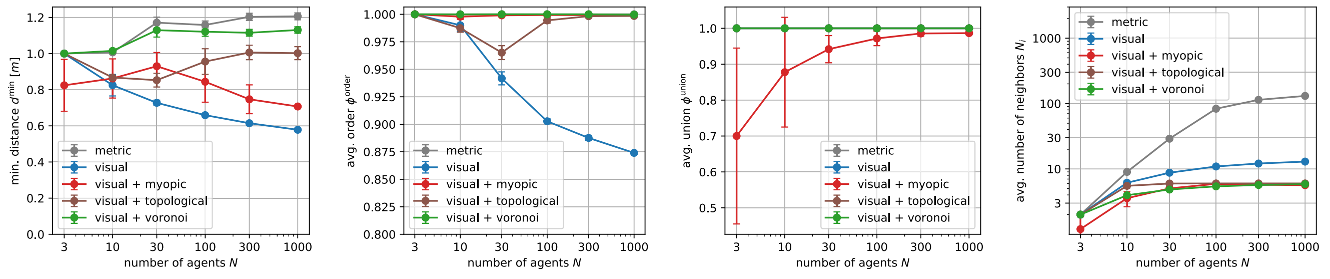
The trajectories of agents using purely *visual* neighbor selection are subject to frequent directional changes (Fig. 5a). As a result, the agents migrate with considerable deviations from the optimal linear trajectory in the migration direction. In particular, the relative positions of the agents within the swarm are not fixed but rather subject to frequent topology switches. For instance, agents that initially belong to the swarm periphery move towards the swarm center (Fig. 5a; blue line) and vice versa.

The topology switches can be explained by considering that an agent within the swarm is exposed to constant changes of its neighbor set (Fig. 6). Small agent displacements result in considerable changes of perspective that cause neighbors to appear and disappear from the visible set (Fig. 6a and 6b: 11 agents appear and 4 disappear, for example). Here, the focal agent is exposed to a total of 32 visibility switches (8 ± 1.22 switches per timestep) over the course of four consecutive seconds of the experiment.

2) ALTERNATIVES TO PURELY VISUAL NEIGHBOR SELECTION

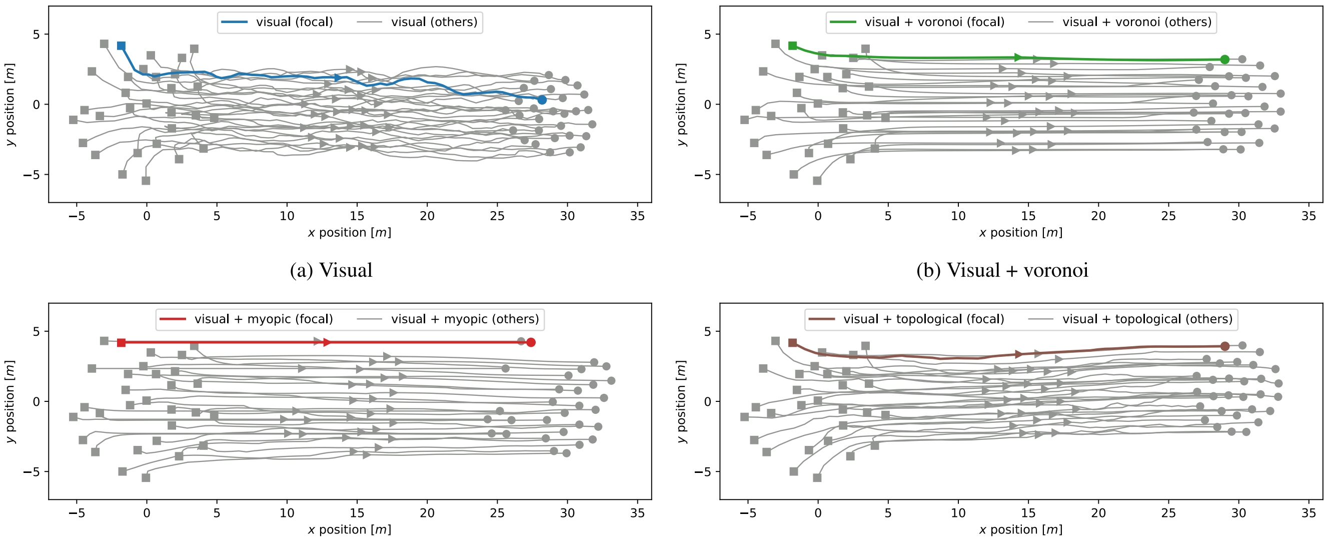
Neighbor selection based on the Voronoi tesselation shows the highest performance of all neighbor selection methods across group sizes. The minimum distance, order, and union metrics show performance comparable to *metric* neighbor selection (Fig. 4a, 4b, and 4c). In particular, the minimum distance is tracked even closer to the reference distance of $d^{\text{ref}} = 1\,\text{m}$ for increasing group size (for 1000 agents: $d^{\text{min}} = 1.13 \pm 0.02\,\text{m}$ for *visual* and $d^{\text{min}} = 1.21 \pm 0.02\,\text{m}$ for *metric*, for example; Fig. 4a). This can be explained by considering that *metric* swarms have a significantly larger number of neighbors compared those based on *visual* + *voronoi* neighbor selection for group sizes $N > 3$ (Fig. 4d). For example, at $N = 1000$ agents, the *metric* neighbor set contains around 22 times the number of agents than it does for *visual* + *topological* neighbor selection (on average 11.2 ± 8.5 times the number of neighbors for all group sizes; Fig. 4d). Recall that the flocking algorithm computes the separation term as a sum of reciprocal distances (Eq. 3). Therefore, each neighbor has an additive contribution towards the repulsion (albeit a very small one for distant agents) that explains the slightly larger distances. The agents are perfectly ordered and cohesive for all group sizes ($\phi^{\text{order}} = 1.0 \pm 0.0$ and $\phi^{\text{union}} = 1.0 \pm 0.0$, respectively; Fig. 4b and 4c). Qualitatively, the paths taken by *visual* + *voronoi* swarms are generally linear and smooth (Fig. 5b). The swarm performs collision-free, ordered, and cohesive collective migration. Switches in the neighbor set do occur but are infrequent and do not lead to unsafe situations or disorder (e.g., changes in neighbor configuration at $x \approx 23\,\text{m}$; Fig. 5b).

Swarms that use *visual* + *myopic* or *visual* + *topological* neighbor selection do not perform as well as those using *visual* + *voronoi* selection for different group sizes. Generally, *visual* + *myopic* swarms exhibit low cohesion and easily fragment into several subgroups (Fig. 4c). Fragmentation occurs because agents that exit the perception radius are usually found within small subgroups or entirely isolated due to their limited perception range (see subgroups and isolated agent; Fig. 5c). The fragmentation phenomenon also skews the minimum distance metric towards lower values with large standard deviations compared to other neighbor selection methods (average of $d^{\text{min}} = 0.82 \pm 0.12\,\text{m}$ across group sizes; Fig. 4a). This occurs because isolated agents are usually



(a) Distance (b) Order (c) Union (d) Neighbors

FIGURE 4. Swarm performance during the collective migration experiment for different neighbor selection methods (Tab. 1) and reference distance $d^{ref} = 1\,\mathrm{m}$. We show the effect of different neighbor selection methods on the (a) minimum nearest neighbor distance d^{min}, (b) average order ϕ^{order}, (c) average union ϕ^{union}, and (d) the average number of neighbors N_i, expressed as a function of the number of agents N (note the logarithmic scale). The neighbors are selected as follows: 1) *metric* selects all agents within the perception radius $r^{max} = 10\,\mathrm{m}$, 2) *visual* selects all visible agents, 3) *visual* + *myopic* selects all visible agents within a smaller radius $r^{max} = 2\,\mathrm{m}$, 4) *visual* + *topological* selects the $n = 6$ topologically closest visible neighbors, and 5) *visual* + *voronoi* selects the neighbors from adjacent Voronoi regions. The Voronoi neighbor selection method scales most predictably with the number of vision-based agents, i.e., distance, order, and union remain quasi-constant as the swarm size increases.



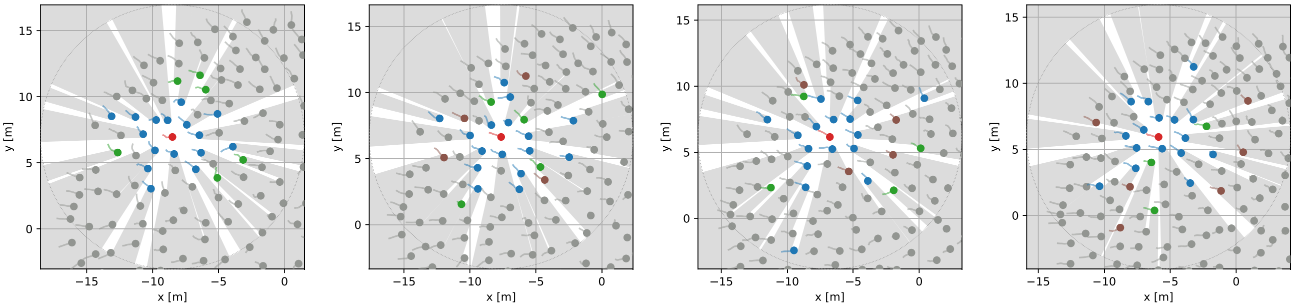
(a) Visual (b) Visual + voronoi

(c) Visual + myopic (d) Visual + topological

FIGURE 5. Example paths taken by a swarm of thirty agents during a single run of the collective migration experiment using the different neighbor selection mechanisms. We use the same random seed to create equal initial conditions and highlight an arbitrary focal agent (colored, thick line) to reveal its motion among the other agents (grey, thin lines). The agents start from their initial positions (solid squares) on the left and migrate along the horizontal axis (solid triangles) to the right side of the virtual arena (solid disks). (a) Visual neighbor selection leads to control discontinuities and disorder; agents frequently change positions inside the swarm. (b) Visual and Voronoi neighbor selection together result in collision-free, ordered, and cohesive migration. (c) Myopic visual interactions also mitigate the discontinuities but lead to fragmentation. (d) Visuo-topological interactions mitigate strong discontinuities but swarms are not well-ordered, especially for peripheral agents.

far away from any other agent (see isolated agent; Fig. 5c). We verified that minimum distances to nearest neighbors are usually well-tracked within subgroups of at least three agents. The union metric is always below $\phi^{union} < 1$ which indicates that fragmentation occurs for all group sizes (Fig. 4c). Cohesion is lowest for small groups and approaches, but never reaches, a value of $\phi^{union} = 1$ that would indicate a single-unit cohesive swarm ($\phi^{union} = 0.7 \pm 0.25$ for $N = 3$, up to $\phi^{union} = 0.98 \pm 0.0$ for $N = 1000$; Fig. 4c). Note that larger groups exhibit higher union performance since the metric is normalized by group size, i.e., larger groups consist of fewer subgroups relative to the overall group size. Swarms with *visual* + *myopic* neighbor selection are effectively ordered ($\phi^{order} = 1.0{\pm}0.0$; Fig. 4b) Qualitatively, apart from fragmentation, larger subgroups tend to have irregular shapes that are less circular compared to other neighbor selection methods (see the largest subgroup; Fig. 5c).

Swarms that use *visual* + *topological* neighbor selection do not exhibit consistent performance accross swarm sizes. Especially for intermediate group sizes of 10, 30, and 100 agents, both minimum distances and order metrics suffer a decrease in performance (Fig. 4a and 7b, respectively). For the respective distances and order metrics, the minimum performance occurs at 30 agents ($d^{min} = 0.85 \pm 0.04\,\mathrm{m}$



(a) $t = 1\,\mathrm{s}$ (b) $t = 2\,\mathrm{s}$ (c) $t = 3\,\mathrm{s}$ (d) $t = 4\,\mathrm{s}$

FIGURE 6. Visual representation of the switching topologies caused by occlusions during a collective migration experiment. We show the perspective of an arbitrary focal agent (central red disk) over the course of four isochronous time steps $t \in \{1\,\mathrm{s}, 2\,\mathrm{s}, 3\,\mathrm{s}, 4\,\mathrm{s}\}$. The focal agent uses *visual* neighbor selection and therefore perceives only agents within its perception radius that are in a direct line of sight (blue disks), whereas occluded agents are invisible (grey disks). We further highlight visibility switches, i.e., when an agent that has been occluded since the previous time step becomes visible (green disks) and when a previously visible agent becomes occluded (brown disks). A total of 32 visibility switches occur over the course of four seconds.

and $\phi^{\text{order}} = 0.97 \pm 0.01$; Fig. 4a and 4b, respectively). We can explain this behavior by considering that agents *always* select the six closest visible neighbors, irrespective of where they are located. Agents that belong to the swarm center tend to have six neighbors that are spaced around them at approximately equal angles from each other. Conversely, agents on the periphery consider only neighbors in one direction which are subject to occlusions. This leads to similar visual switching topologies as for the purely *visual* neighbor selection, albeit less severe since even the most distant nearest neighbor for $n = 6$ is usually in close proximity. The effect of occlusions is mostly mitigated for larger swarm sizes $N > 100$ since a smaller proportion of agents is located on the periphery relative to the swarm center. We do not observe fragmentation with *visual* + *topological* neighbor selection for any group size ($\phi^{\text{union}} = 1.0 \pm 0.0$; Fig. 4c). Qualitatively, *visual* + *topological* interactions generate paths that are not perfectly straight (Fig. 5d). We also observe swarms that exhibit rotations, as well as ones that periodically switch between a set of recurring configurations.

B. PERFORMANCE ACROSS SWARM DENSITIES

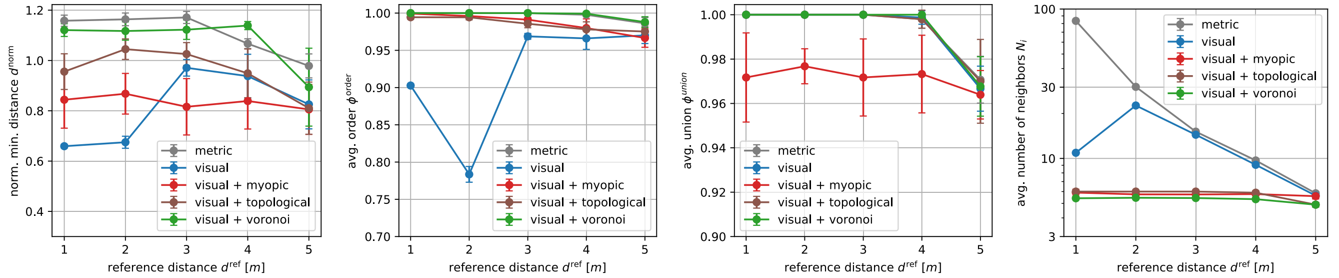
We evaluate the swarm performance for all neighbor selection methods and for five levels of increasing inter-agent distances $d^{\text{ref}} \in \{1\,\mathrm{m}, 2\,\mathrm{m}, 3\,\mathrm{m}, 4\,\mathrm{m}, 5\,\mathrm{m}\}$. We let $N = 100$ to fix the group size and to enable a direct comparison between agent number densities. We define the normalized minimum nearest neighbor distance as $d^{\text{norm}} = d^{\text{min}}/d^{\text{ref}}$ to make the minimum distances more easily comparable for different agent densities.

1) VISUAL NEIGHBOR SELECTION

Purely *visual* neighbor selection does not show consistent performance for different swarm densities. The performance penalty in distance and order is especially severe for agents in high-density configurations with small reference distances (Fig. 7a and 7b, respectively). The normalized distance is much lower than the desired reference of $d^{\text{norm}} \geq 1$ and has its minimum for $d^{\text{ref}} \in \{1\,\mathrm{m}, 2\,\mathrm{m}\}$ ($d^{\text{norm}} = 0.66 \pm 0.01$ and $d^{\text{norm}} = 0.67 \pm 0.02$, respectively; Fig. 7a). For larger reference distances, $d^{\text{ref}} \in \{3\,\mathrm{m}, 4\,\mathrm{m}\}$, the normalized distance stabilizes again to larger values ($d^{\text{norm}} = 0.97 \pm 0.03$ and $d^{\text{norm}} = 0.94 \pm 0.09$, respectively; Fig. 7a) Note that the minimum distance, order, and union metrics for large reference distances $d^{\text{ref}} = 5\,\mathrm{m}$ decrease for all neighbor selection methods. A reference distance of $d^{\text{ref}} = r^{\text{max}}/2 = 5\,\mathrm{m}$ effectively renders all neighbor selection methods *myopic* and fragmentation starts to occur. The union metric indicates that this is indeed the case for $d^{\text{ref}} = 5\,\mathrm{m}$ since all neighbor selection methods show comparable mean performance to *myopic* swarms (average of all neighbor selection methods $\phi^{\text{union}} = 0.97 \pm 0.01$; Fig. 7c). The order metric reaches its minimum at $d^{\text{ref}} = 2\,\mathrm{m}$ ($\phi^{\text{order}} = 0.78 \pm 0.01$; Fig. 7b). The minimum order coincides with the maximum of the average number of visible neighbors at $d^{\text{ref}} = 2\,\mathrm{m}$ ($N_i = 22.62 \pm 0.04$). This indicates that order follows an inverse relationship with the number of visible neighbors: if more agents are visible, the likelihood of visual topology switches that lead to disorder increases (Fig. 6). The neighbor graph also highlights that the effect of occlusions is maximized at intermediate densities. At high densities, the nearest neighbors occlude most agents in all directions (87% occluded for $d^{\text{ref}} = 1\,\mathrm{m}$; Fig. 7d). Conversely, the effect of occlusions diminishes at lower densities since the agents are not large enough to break the line of sight (5% occluded for $d^{\text{ref}} = 3\,\mathrm{m}$, for example; Fig. 7d).

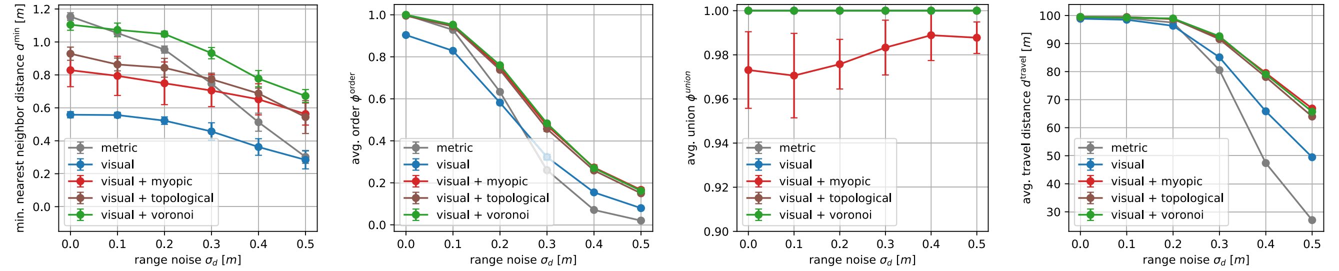
2) ALTERNATIVES TO PURELY VISUAL NEIGHBOR SELECTION

The Voronoi-based neighbor selection provides the highest and most consistent performance across different group densities. The distance, order, and union metrics remain stable for all but the lowest density level ($d^{\text{ref}} = 5\,\mathrm{m}$) at which interactions are rendered *myopic* (Fig. 7a, 7b, and 7c;



(a) Distance (b) Order (c) Union (d) Neighbors

FIGURE 7. Swarm performance during the collective migration experiment for different neighbor selection methods (Tab. 1) and group size $N = 100$. We show the effect of different neighbor selection methods on the (a) normalized minimum nearest neighbor distance d^{norm}, (b) average order ϕ^{order}, (c) average union ϕ^{union}, and (d) average number of neighbors N_i, expressed as a function of the reference distance d^{ref}. With the exception of *myopic* conditions (at $d^{\text{ref}} = 5\,\text{m}$), the Voronoi neighbor selection method scales most predictably with the density of the vision-based swarm and the performance remains quasi-constant as the reference distance increases.



(a) Distance (b) Order (c) Union (d) Travel

FIGURE 8. Swarm performance during the collective migration experiment for different neighbor selection methods (Tab. 1) with group size $N = 100$ and $d^{\text{ref}} = 1\,\text{m}$. We show the effect of different neighbor selection methods on the (a) minimum nearest neighbor distance d^{min}, (b) average order ϕ^{order}, (c) average union ϕ^{union}, and (d) average travel distance d^{travel}, expressed as a function of the range noise σ_d. Overall, the Voronoi-based neighbor selection method shows comparable (in terms of order and travel distance) or higher performance (in terms of minimum distance and union) than the vision-based alternatives for increasing noise levels.

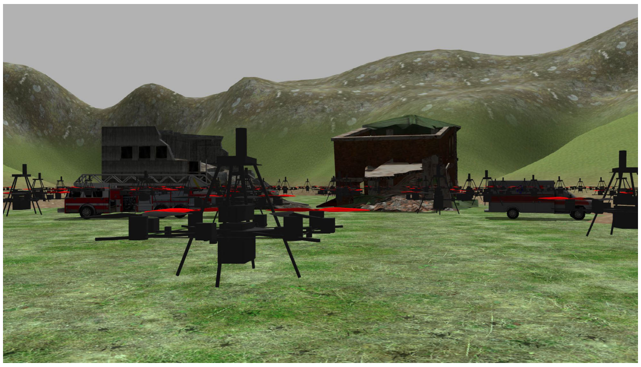


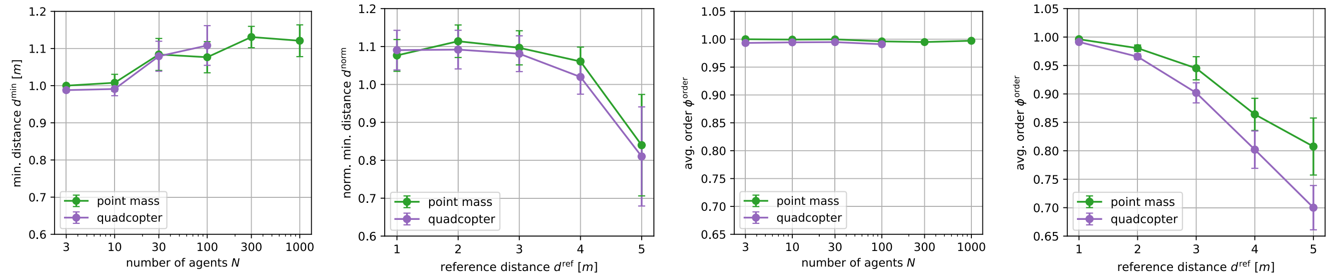
FIGURE 9. Screenshot of a collective search and rescue mission with a swarm of one hundred quadcopters in the Gazebo simulator. During the mission, the quadcopters take off from the ground and navigate towards a fictitious disaster scenario.

Sec. IV-B1 for discussion of *myopic* interactions). The normalized distance, order, and union remain stable for high and intermediate swarm densities (average $d^{\text{norm}} = 1.12 \pm 0.03\,\text{m}$, $\phi^{\text{order}} = 1.0 \pm 0.0$, and $\phi^{\text{union}} = 1.0 \pm 0.0$; Fig. 7a, 7b, 7c, respectively).

Swarms with *visual* + *myopic* and *visual* + *topological* interactions perform comparatively poorly to *visual* + *voronoi* neighbor across group densities. The *visual* + *myopic* neighbor selection method shows consistently low performance in terms of distance and union metrics (Fig. 7a and 7c). Myopic interactions effectively reduce the negative impact of occlusions. However, they also induce low distances and fragmentation (average $d^{\text{norm}} = 0.84 \pm 0.02$ and $\phi^{\text{union}} = 0.97 \pm 0.01$ across reference distances; Fig. 7a and 7c, respectively). Swarms with *visual* + *topological* interactions can avoid the fragmentation issues but their minimum distances fluctuate for different densities (e.g., $d^{\text{norm}} = 1.05 \pm 0.04$ for $d^{\text{ref}} = 2\,\text{m}$ and $d^{\text{norm}} = 0.95 \pm 0.10$ for $d^{\text{ref}} = 4\,\text{m}$; Fig. 7a).

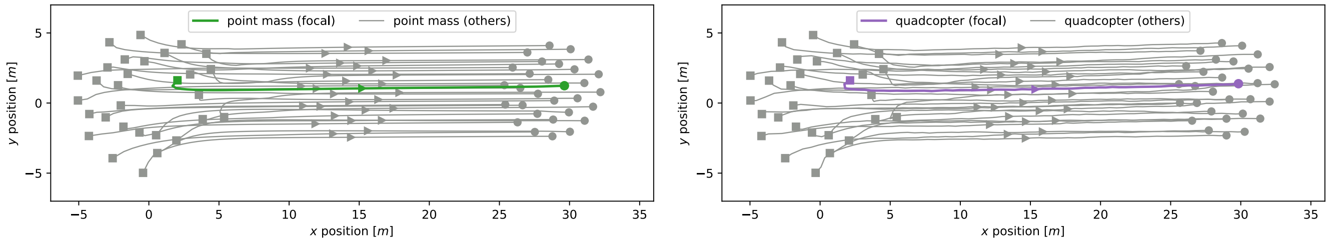
C. PERFORMANCE ACROSS NOISE LEVELS

We assess the performance of the swarm for all neighbor selection methods and six levels of increasing range noise $\sigma_d \in \{0.0\,\text{m}, 0.1\,\text{m}, 0.2\,\text{m}, 0.3\,\text{m}, 0.4\,\text{m}, 0.5\,\text{m}\}$, which corresponds directly to a percentage at the chosen reference distance $d^{\text{ref}} = 1\,\text{m}$. We vary the range standard deviation in 0.1 m-increments while keeping both group size $N = 100$ and reference distance $d^{\text{ref}} = 1\,\text{m}$ fixed to compare



(a) Distance vs. group size (b) Distance vs. density (c) Order vs. group size (d) Order vs. density

FIGURE 10. Quantitative results from the collective migration experiment comparing point mass dynamics (green) and with quadcopter dynamics and realistic noise (purple). We show the effect of the simulation realism on the (a, b) minimum (normalized) nearest neighbor distances d^{min} (d^{norm}) and (c, d) average order ϕ^{order} metrics. The metrics are shown as a function of (a, c) the number of agents *N* (note the logarithmic scale) and (b, d) the reference distance d^{ref}. The metrics follow very similar trends under point mass and quadcopter dynamics with realistic noise. For quadcopter dynamics, we limit our analysis to one hundred agents since simulations with larger group sizes proved to be too unreliable for statistical analysis.



(a) Point mass dynamics without noise (b) Quadcopter dynamics with noise

FIGURE 11. Example paths taken by a swarm of thirty agents during a single run of the collective migration experiment using (a) point mass dynamics without noise and (b) quadcopter dynamics with realistic noise. We use the same random seed to create equal initial conditions and highlight an arbitrary focal agent (colored, thick line) to reveal its motion among the other agents (grey, thin lines). The agents start from their initial positions (solid squares) on the left and migrate along the horizontal axis (solid triangles) to the right side of the virtual arena (solid disks). Apart from the effect of noise, there is no discernable qualitative difference between the point mass and quadcopter swarms.

how well the neighbor selection methods perform at different noise levels.

1) VISUAL NEIGHBOR SELECTION

Purely *visual* neighbor selection shows the overall lowest performance in terms of minimum nearest neighbor distance for all noise levels. Collisions between agents start to occur with a noise level of $\sigma_d = 0.3$ m ($d^{min} = 0.46 \pm 0.05$ m; Fig. 8a). Purely *visual* neighbor selection also exhibits the lowest average order in comparison to the other *visual* neighbor selection methods (*myopic*, *topological*, and *voronoi*). The average order for purely *visual* neighbor selection follows the same trend as the other *visual* neighbor selection methods as noise increases, however with a consistently lower average order (difference of about $\phi^{order} = 0.1$; Fig. 8b). Purely *visual* swarms do not separate into subflocks even as noise increases, as evidenced by their perfect union score ($\phi^{union} = 1.0 \pm 0.0$; Fig. 8c). On average, purely *visual* swarms travel only roughly half as far when subjected to 50% range noise compared to when they operate without noise ($d^{travel} = 49.47 \pm 0.15$ m at $\sigma_d = 0.5$ m vs. $d^{travel} = 98.87 \pm 0.29$ m at $\sigma_d = 0.0$ m; Fig. 8d).

2) ALTERNATIVES TO VISUAL NEIGHBOR SELECTION

Overall, the *visual* + *voronoi* neighbor selection method shows similar or higher performance scores than the other vision-based alternatives (namely, *visual* + *myopic* and *visual* + *topological* interactions) across the evaluated metrics and noise levels. Regarding the minimum nearest neighbor distance, *visual* + *voronoi* neighbor selection outperforms the vision-based alternatives for all noise levels (highest score $d^{min} = 1.10 \pm 0.03$ m for $\sigma_d = 0.0$ m and lowest score $d^{min} = 0.67 \pm 0.04$ m for $\sigma_d = 0.5$ m; Fig. 8a). Generally, the minimum nearest neighbor distance of the vision-based neighbor selection methods show a similar downward trend for increasing noise levels (Fig. 8a). For comparison, *metric* neighbor selection is much more sensitive to increasing noise levels and has the largest difference between performance scores (maximum $d^{min} = 1.15 \pm 0.02$ m for $\sigma_d = 0.0$ m and minimum $d^{min} = 0.30 \pm 0.04$ m for $\sigma_d = 0.5$ m; Fig. 8a). For the order and travel distance metrics, *visual* + *voronoi* interactions perform comparable to the other vision-based alternatives (Fig. 8b and Fig. 8d). In terms of union metric, only *visual* + *myopic* interactions break the swarm into subflocks (Fig. 8c). Interestingly, the union metric also increases with higher noise levels which allow separate subflocks to reunite occasionally (minimum $\phi^{union} = 0.97 \pm 0.02$ at $\sigma_d = 0.0$ m vs. maximum $\phi^{union} = 0.99 \pm 0.01$ at $\sigma_d = 0.5$ m; Fig. 8c).

D. VALIDATION IN REALISTIC CONDITIONS

We finally assess the performance of the most promising *visual* + *voronoi* neighbor selection method in more realistic conditions. This is done to evaluate whether the performance transfers to agents with quadcopter dynamics and more realistic sensor noise. Analogous to the previous experiments, we vary the reference distance $d^{\text{ref}} = \{1\,\text{m}, 2\,\text{m}, 3\,\text{m}, 4\,\text{m}, 5\,\text{m}\}$ while keeping the number of agents $N = 100$ fixed to show the effect of agent number density on the flocking performance. Similarly, we vary the number of agents $N \in \{3, 10, 30, 100\}$ ($N \leq 100$ due to the limitations of the physics simulation; Sec. III-C) while keeping the reference distance $d^{\text{ref}} = 1\,\text{m}$ constant to show the effect of group size on the performance metrics. We replace the single-integrator dynamics (Eq. 1) with a cascaded PID controller [51] that uses the velocity commands from the flocking algorithm as inputs (Eq. 2). We further set the range and bearing noise to $\sigma_d = 0.05\,\text{m}$ and $\sigma_\beta = 1°$, respectively. The specific values are informed by our previous experiments in indoor [14] and outdoor environments [18] and resemble estimates from visual relative localization using object detection with a multi-target state tracker [52] that was specifically tuned for the operating conditions. The exact noise values may be higher if raw observations are used and depend on many factors such as detector performance, camera resolution, and background clutter.

The *visual + voronoi* neighbor selection method shows comparable performance with point mass agents and quadcopters operating with realistic sensor noise. The swarm performance generally degrades more with increasing reference distances than it does for increasing group size, regardless of the simulation realism. We omit an analysis of the union since the swarms remained cohesive as a single unit during all experiments without exception ($\phi^{\text{union}} = 1.0 \pm 0.0$). We further omit the neighbor statistics since we did not observe any discernable differences. The only noticeable difference between *point mass* and *quadcopter* simulations is the divergence of the average order for decreasing density ($\phi^{\text{order}} = 0.81 \pm 0.05$ for *point mass* and $\phi^{\text{order}} = 0.70 \pm 0.04$ *quadcopter*; Fig. 10d). This difference can largely be attributed to the range noise that increases linearly with distance (Eq. 10). The effect of the noise for *quadcopter* dynamics can also be observed in the slightly lower normalized distances compared to *point mass* dynamics (Fig. 10a) Interestingly, the more realistic simulation also results in slightly larger minimum distances for 100 agents than would be expected with decreases due to noise ($d^{\text{min}} = 1.07 \pm 0.04\,\text{m}$ for *point mass* and $d^{\text{min}} = 1.11 \pm 0.05\,\text{m}$ for *quadcopter*; Fig. 10a). However, these effects are too small to be significant and could have occurred due to chance.

V. CONCLUSION

Methods for multi-agent coordination often make unrealistic assumptions about the information that is available to the individual agent. One of the most pervasive simplifying assumptions is that vision-based agents can sense the state of *all* surrounding neighbors within a metric perception radius, even if they are obstructed by closer ones. Here, we break this common assumption and construct a simple yet realistic model of visibility that selects neighbors only if 1) they appear large enough in the field of view, and 2) are not occluded by other agents. Extensive flocking simulations with the visual occlusion model show that perfectly ordered metric-based swarms become disordered and unsafe when agents react to all of their visible neighbors. These adverse effects can be attributed to small perspective changes that continuously influence the set of visible neighbors, thus causing the agents to move in reaction to the new neighbor configuration. We show that this interplay between visibility constraints and collective motion can lead to severe instabilities for vision-based swarms, especially for large numbers of agents and high swarm densities.

Selecting a subset of visible neighbors from adjacent Voronoi regions significantly improves the swarm performance (i.e., collision avoidance, velocity alignment, group cohesion, and travel distance) across group sizes, agent number densities, and noise levels. Controlled experiments with subsets of the visual neighbors show that Voronoi-based interactions are a more effective countermeasure against occlusions than metric and topological ones. The main drawback of metric and topological neighbor selection methods is their dependence on specific parameters, namely the perception range and the number of nearest neighbors, respectively. Choosing favorable values for these parameters that provide high performance at all group sizes and densities may be impossible for vision-based swarms. In particular, swarms that select too many neighbors suffer from the adverse effects of occlusions and selecting too few neighbors inevitably leads to fragmentation. Voronoi-based interactions provide an elegant solution to this problem since they are both parameter-free and spatially balanced [30].

The occlusion model presented here is undoubtedly useful but it neglects two important aspects of vision-based relative localization: errors due to misdetections (false positive and false negatives) and partial occlusions. False positives (i.e., detecting an agent that is not there) and false negatives (i.e., not detecting an agent that is defacto there) inevitably occur in real-world conditions but are notoriously difficult to model. Multi-target filtering algorithms can alleviate errors due to sensing noise and false positive detections to some extent but are largely ineffective against false negatives [18]. Modeling partial visual occlusions is equally challenging; agents that occlude others with a given overlap may themselves be — possibly recursively— occluded by other agents at different locations. Whether multiple partially occluded agents should be detected as a single agent at a closer distance is another modeling choice to consider. The main difficulty is that the distribution of these errors depends not only on the robot's physical appearance and the error distribution of the detection algorithm but also on environmental conditions such as background clutter and lighting conditions. We believe that modeling these factors based on first principles is of limited use due to many arbitrary modeling choices. Future work should therefore systematically characterize misdetections and partial occlusions in a more realistic setting with vision-based detectors that localize physical robots in real images. This characterization could then inform many modeling choices, e.g., temporal and spatial distribution of false positives and overlap thresholds for partial occlusions.

We argue that occlusions should not be neglected when designing algorithms for vision-based swarms. We consider the simple occlusion model presented here (Eq. 6) as a useful drop-in replacement for vision-based flocking algorithms that would otherwise default to purely metric interactions (Eq. 5). Simple agent-based simulations can thus prevent significant hardware damage by considering occlusions early in the algorithm design and before they are implemented on real robots. The validation presented here is specifically geared towards drones but we expect the results to translate well to other types of vision-based robots.

ACKNOWLEDGMENT

The authors would like to thank Vivek Ramachandran, Enrico Ajanic, and Valentin Wüest for the helpful discussions and suggestions.

REFERENCES

[1] D. Floreano and R. J. Wood, ''Science, technology and the future of small autonomous drones,'' *Nature*, vol. 521, pp. 460–466, May 2015.

[2] S.-J. Chung, A. A. Paranjape, P. Dames, S. Shen, and V. Kumar, ''A survey on aerial swarm robotics,'' *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 837–855, Aug. 2018.

[3] M. Coppola, K. N. McGuire, C. De Wagter, and G. C. H. E. de Croon, ''A survey on swarming with micro air vehicles: Fundamental challenges and constraints,'' *Frontiers Robot. AI*, vol. 7, p. 18, Feb. 2020.

[4] S. Hauert, S. Leven, M. Varga, F. Ruini, A. Cangelosi, J.-C. Zufferey, and D. Floreano, ''Reynolds flocking in reality with fixed-wing robots: Communication range vs. maximum turning rate,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2011, pp. 5015–5020.

[5] E. Soria, F. Schiano, and D. Floreano, ''Predictive control of aerial swarms in cluttered environments,'' *Nat. Mach. Intell.*, vol. 3, pp. 545–554, May 2021.

[6] T. Cieslewski, S. Choudhary, and D. Scaramuzza, ''Data-efficient decentralized visual SLAM,'' in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2018, pp. 2466–2473.

[7] G. Vásárhelyi, C. Virágh, G. Somorjai, N. Tarcai, T. Szörenyi, T. Nepusz, and T. Vicsek, ''Outdoor flocking and formation flight with autonomous aerial robots,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 3866–3873.

[8] G. Vásárhelyi, C. Virágh, G. Somorjai, T. Nepusz, A. E. Eiben, and T. Vicsek, ''Optimized flocking of autonomous drones in confined environments,'' *Sci. Robot.*, vol. 3, no. 20, Jul. 2018, Art. no. eaat3536.

[9] M. Coppola, K. N. McGuire, K. Y. W. Scheper, and G. C. H. E. de Croon, ''On-board communication-based relative localization for collision avoidance in micro air vehicle teams,'' *Auto. Robots*, vol. 42, no. 8, pp. 1787–1805, Dec. 2018.

[10] J. P. Queralta, C. M. Almansa, F. Schiano, D. Floreano, and T. Westerlund, ''UWB-based system for UAV localization in GNSS-denied environments: Characterization and dataset,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4521–4528.

[11] K. N. McGuire, C. De Wagter, K. Tuyls, H. J. Kappen, and G. C. H. E. de Croon, ''Minimal navigation solution for a swarm of tiny flying robots to explore an unknown environment,'' *Sci. Robot.*, vol. 4, no. 35, Oct. 2019, Art. no. eaaw9710.

[12] T. Qin, P. Li, and S. Shen, ''VINS-mono: A robust and versatile monocular visual-inertial state estimator,'' *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[13] P. M. Wyder, Y.-S. Chen, A. J. Lasrado, R. J. Pelles, R. Kwiatkowski, E. O. A. Comas, R. Kennedy, A. Mangla, Z. Huang, X. Hu, Z. Xiong, T. Aharoni, T.-C. Chuang, and H. Lipson, ''Autonomous drone hunter operating by deep learning and all-onboard computations in GPS-denied environments,'' *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0225092.

[14] F. Schilling, J. Lecoeur, F. Schiano, and D. Floreano, ''Learning vision-based flight in drone swarms by imitation,'' *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4523–4530, Oct. 2019.

[15] M. Vrba and M. Saska, ''Marker-less micro aerial vehicle detection and localization using convolutional neural networks,'' *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2459–2466, Apr. 2020.

[16] M. Saska, T. Baca, J. Thomas, J. Chudoba, L. Preucil, T. Krajnik, J. Faigl, G. Loianno, and V. Kumar, ''System for deployment of groups of unmanned micro aerial vehicles in GPS-denied environments using onboard visual relative localization,'' *Auton. Robots*, vol. 41, no. 4, pp. 919–944, 2017.

[17] P. Petráček, V. Walter, T. Báča, and M. Saska, ''Bio-inspired compact swarms of unmanned aerial vehicles without communication and external localization,'' *Bioinspiration Biomimetics*, vol. 16, no. 2, Mar. 2021, Art. no. 026009.

[18] F. Schilling, F. Schiano, and D. Floreano, ''Vision-based drone flocking in outdoor environments,'' *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2954–2961, Apr. 2021.

[19] J. Chen, M. Gauci, W. Li, A. Kolling, and R. Groß, ''Occlusion-based cooperative transport with a swarm of miniature mobile robots,'' *IEEE Trans. Robot.*, vol. 31, no. 2, pp. 307–321, Apr. 2015.

[20] J. Hu, A. E. Turgut, T. Krajnik, B. Lennox, and F. Arvin, ''Occlusion-based coordination protocol design for autonomous robotic shepherding tasks,'' *IEEE Trans. Cognit. Develop. Syst.*, early access, Aug. 21, 2021, doi: 10.1109/TCDS.2020.3018549.

[21] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, ''Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study,'' *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1232–1237, 2008.

[22] A. Strandburg-Peshkin, C. R. Twomey, N. W. F. Bode, A. B. Kao, Y. Katz, C. C. Ioannou, S. B. Rosenthal, C. J. Torney, H. S. Wu, S. A. Levin, and I. D. Couzin, ''Visual sensory networks and effective information transfer in animal groups,'' *Current Biol.*, vol. 23, no. 17, pp. R709–R711, Sep. 2013.

[23] S. B. Rosenthal, C. R. Twomey, A. T. Hartnett, H. S. Wu, and I. D. Couzin, ''Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion,'' *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 15, pp. 4690–4695, Apr. 2015.

[24] J. D. Davidson, M. M. G. Sosna, C. R. Twomey, V. H. Sridhar, S. P. Leblanc, and I. D. Couzin, ''Collective detection based on visual information in animal groups,'' *J. Roy. Soc. Interface*, vol. 18, no. 180, Jul. 2021, Art. no. 20210142.

[25] T. Vicsek and A. Zafeiris, ''Collective motion,'' *Phys. Rep.*, vol. 517, no. 3, pp. 71–140, 2012.

[26] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, ''Collective memory and spatial sorting in animal groups,'' *J. Theor. Biol.*, vol. 218, no. 1, pp. 1–11, Sep. 2002.

[27] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin, ''Effective leadership and decision-making in animal groups on the move,'' *Nature*, vol. 433, no. 7025, pp. 513–516, 2005.

[28] A. E. Turgut, H. Çelikkana, F. Gökçe, and E. Şahin, ''Self-organized flocking in mobile robot swarms,'' *Swarm Intell.*, vol. 2, nos. 2–4, pp. 97–120, 2008.

[29] R. Olfati-Saber, ''Flocking for multi-agent dynamic systems: Algorithms and theory,'' *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 401–420, Mar. 2006.

[30] M. Camperi, A. Cavagna, I. Giardina, G. Parisi, and E. Silvestri, ''Spatially balanced topological interaction grants optimal cohesion in flocking models,'' *Interface Focus*, vol. 2, no. 6, pp. 715–725, Dec. 2012.

[31] M. Lindhe, P. Ogren, and K. H. Johansson, ''Flocking with obstacle avoidance: A new distributed coordination algorithm based on Voronoi partitions,'' in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2005, pp. 1785–1790.

[32] J. Holland and S. K. Semwal, ''Flocking boids with geometric vision, perception and recognition,'' in *Proc. Int. Conf. Central Eur. Comput. Graph., Vis. Comput. Vis. (WSCG)*, 2009. [Online]. Available: http://wscg.zcu.cz/WSCG2009/wscg2009.htm

[33] E. Soria, F. Schiano, and D. Floreano, ''The influence of limited visual sensing on the Reynolds flocking algorithm,'' in *Proc. 3rd IEEE Int. Conf. Robotic Comput. (IRC)*, Feb. 2019, pp. 138–145.

[34] G. R. Martin, ''Visual fields and their functions in birds,'' *J. Ornithol.*, vol. 148, no. S2, pp. 547–562, Dec. 2007.

[35] H. Kunz and C. K. Hemelrijk, "Simulations of the social organization of large schools of fish whose perception is obstructed," *Appl. Animal Behav. Sci.*, vol. 138, nos. 3–4, pp. 142–151, May 2012.
[36] D. J. G. Pearce, A. M. Miller, G. Rowlands, and M. S. Turner, "Role of projection in the control of bird flocks," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 29, pp. 10422–10426, Jul. 2014.
[37] R. Bastien and P. Romanczuk, "A model of collective behavior based purely on vision," *Sci. Adv.*, vol. 6, no. 6, Feb. 2020, Art. no. eaay0792.
[38] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 25–34, Aug. 1987.
[39] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, "Stable flocking of mobile agents—Part II: Dynamic topology," in *Proc. 42nd IEEE Conf. Decis. Control*, vol. 2, Dec. 2003, pp. 2016–2021.
[40] B. T. Fine and D. A. Shell, "Unifying microscopic flocking motion models for virtual, robotic, and biological flock members," *Auto. Robots*, vol. 35, nos. 2–3, pp. 195–219, Oct. 2013.
[41] Y. Shang and R. Bouffanais, "Influence of the number of topologically interacting neighbors on swarm dynamics," *Sci. Rep.*, vol. 4, no. 1, p. 4184, May 2015.
[42] A. Okabe, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Probability and Statistics), 2nd ed. Hoboken, NJ, USA: Wiley, 2000.
[43] S. Roelofsen, D. Gillet, and A. Martinoli, "Reciprocal collision avoidance for quadrotors using on-board visual detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4810–4817.
[44] D. Dias, R. Ventura, P. Lima, and A. Martinoli, "On-board vision-based 3D relative localization system for multiple quadrotors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 1181–1187.
[45] V. Walter, N. Staub, A. Franchi, and M. Saska, "UVDAR system for visual relative localization with application to leader–follower formations of multirotor UAVs," *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, pp. 2637–2644, Jul. 2019.
[46] K. R. Sapkota, S. Roelofsen, A. Rozantsev, V. Lepetit, D. Gillet, P. Fua, and A. Martinoli, "Vision-based unmanned aerial vehicle detection and tracking for sense and avoid systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 1556–1561.
[47] A. Rozantsev, V. Lepetit, and P. Fua, "Detecting flying objects using a single moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 879–892, May 2017.
[48] R. Opromolla, G. Fasano, and D. Accardo, "A vision-based approach to UAV detection and tracking in cooperative applications," *Sensors*, vol. 18, no. 10, p. 3391, Oct. 2018.
[49] M. Vrba, D. Hert, and M. Saska, "Onboard marker-less detection and localization of non-cooperating drones for their safe interception by an autonomous aerial system," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3402–3409, Oct. 2019.
[50] G. Dell'Ariccia, G. Dell'Omo, D. P. Wolfer, and H.-P. Lipp, "Flock flying improves pigeons' homing: GPS track analysis of individual flyers versus small groups," *Animal Behav.*, vol. 76, no. 4, pp. 1165–1172, Oct. 2008.
[51] L. Meier, D. Honegger, and M. Pollefeys, "PX4: A node-based multithreaded open source robotics framework for deeply embedded platforms," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 6235–6240.
[52] B. N. Vo and W. K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.

FABIAN SCHILLING received the M.S. degree in computer science from the KTH Royal Institute of Technology, in 2017. He is currently pursuing the Ph.D. degree in robotics, control, and autonomous systems with the Swiss Federal Institute of Technology Lausanne (EPFL). His research interests include computer vision, machine learning, and robotics.

ENRICA SORIA received the M.S. degree in mathematical engineering from the Polytechnic University of Turin, in 2016. She is currently pursuing the Ph.D. degree in robotics, control, and autonomous systems with the Swiss Federal Institute of Technology Lausanne (EPFL). Her research interests include mathematical modeling, control, and robotics.

DARIO FLOREANO (Senior Member, IEEE) received the M.A. degree in vision, the M.S. degree in neural computation, and the Ph.D. degree in robotics.

He has held research positions at Sony Computer Science Laboratory at Caltech/JPL and at Harvard University. He is currently the Director of the Laboratory of Intelligent Systems, Swiss Federal Institute of Technology Lausanne (EPFL). Since 2010, he has been the Founding Director of the Swiss National Center of Competence in Robotics, a research program that brings together more than 20 labs across Switzerland. His main research interests include robotics and AI at the convergence of biology and engineering. He made pioneering contributions to the fields of evolutionary robotics, aerial robotics, and soft robotics. He served in numerous advisory boards and committees, including the Future and Emerging Technologies Division of the European Commission, the World Economic Forum Agenda Council, the International Society of Artificial Life, the International Neural Network Society, and in the editorial committee of several scientific journals. In addition, he helped to spin off two drone companies (senseFly.com and Flyability.com) and a non-for-profit portal on robotics and AI (RoboHub.org).