Fine-tuned thin-plate spline motion model for manipulating social information in paper-wasp colonies

Kacy Hatfield
khatfie2@asu.edu

Akuadasuo Ezenyilimba
aezenyi1@asu.edu

Nitin Verma
nitinv@asu.edu

Juan José García Mesa
jgarc111@asu.edu

So Eun Moon
soemoon@umich.edu

Elizabeth Tibbetts
tibbetts@umich.edu

Pavan Turaga
pturaga@asu.edu

Theodore P. Pavlic
tpavlic@asu.edu

Abstract

Several species of Polistes *paper wasp are well known for their social hierarchies and the ability for individual wasps to modulate their social behaviors based on recognizable facial features of other wasps. For example, wasps that observe an aggressive social interaction between two other wasps will later behave differently toward the winner and loser of that interaction. Being able to alter the physical appearance of wasps (e.g., with paint) has allowed for testing hypothetical roles of individual recognition in hierarchy formation, which is how researchers know that wasps are attending to faces specifically. However, these physical methods are limited in their scope. Social insects who respond to visual stimuli from other insects have been shown to give the same responses to playbacks of video recordings of those stimuli, which suggests that there may be a role for generative methods in social-insect research. Being able to computationally change the faces of individual wasps in a video recording of wasp social interactions would greatly expand the experimental toolbox of the behavioral researcher. Toward this end, we evaluate the use of an existing annotation-free model for image animation by motion transfer, the thin-plate spline motion model, for creating realistic videos that depict the face of a paper wasp performing behaviors recorded by another. Not needing to pre-define important landmarks is a strength of this method for this application space, but we find that "deep faking wasps" poses unique and non-trivial problems that still need to be solved before off-the-shelf motion transfer models can be used in the insect behavioral laboratory*[1].

[1]This work was supported in part by the National Science Foundation under Grants OAC-2230108, IIS-2323086, and EF-2319438.

1. Introduction

Over the past two decades, paper wasps in the genus *Polistes* have emerged as a promising animal model for studying the role of perceptual systems in the evolution and maintenance of hierarchical social systems [15, 17–19]. For example, colonies of both the northern paper wasp *P. fuscatus* and the European paper wasp *P. dominula* have individuals with a high level of variation in their facial markings (e.g., Figure 1) that play important but very different roles in their reproductive hierarchies. In *P. dominula*, geometric aspects of these facial features are reliable signals of fighting ability and can predict the position of a wasp in her hierarchy; however, *P. fuscatus* facial variation does not correlate with resource holding potential, and individual wasps behave more like primates in that they remember and respond to holistic facial patterns peculiar to each individual in the colony [10, 18]. The primate-like level of social complexity in paper wasps combined with the relatively high degree of empirical tractability makes them an attractive model system for studying animal social behavior [16].

The differences in how *P. dominula* and *P. fuscatus* use facial-pattern information was discovered using experiments that altered the physical appearance of real wasps (e.g., by using paint). Effective as this method has been, it still has significant limitations. For example, "social eavesdropping" has been demonstrated in *P. fuscatus*; an individual who, from a distance, observes an aggressive interaction between two other individuals will remember the face of the winner and act more submissively in later interactions with her [17]. It could be useful to vary the faces of wasps engaged in a single interaction and then study the downstream social effects on the behavior of wasps who observe different types of social interactions. However, paint-based methods do not allow for reproducing the exact same social interaction with only changes in face. Other social insects have been shown to be responsive to video playback of behaviors of other insects [e.g., 2], which suggests that generative video models may be an effective tool in future

(a) *P. dominula* (b) *P. fuscatus*

Figure 1. Examples of variation in the faces of two species of *Polistes* paper wasp. In the European paper wasp *P. dominula* (a), geometric features of the dark spot in the middle of the clypeus at the front of the head are strong predictors of the social dominance of the individual; however, in the northern paper wasp *P. fuscatus* (b), faces are used to individually recognize an individual and recall a history of learned information about it [18].

social-insect laboratories. Ideally, computational methods would allow for recording several prototypical social interactions and then swapping out faces as is convenient for the experimental design. Such a flexible generative video vocabulary of social interactions would allow for unprecedented experimental control of the injection of social information into a functioning animal group.

In this paper, we explore the use of motion-transfer models to recreate recorded behaviors of paper wasps with different faces. As we seek a method that will be generally applicable across a wide taxonomic range of paper wasps (if not insects more broadly), we excluded from consideration models that require labor-intensive manual marking and pre-specification of important landmarks. Instead, the core of our approach lies in fine-tuning an existing thin-plate spline (TPS) motion model from Zhao and Zhang [22] to effectively produce identity swaps for insect studies without the need for manual labels. The TPS motion model offers a significant advancement in the field of motion transfer by using an automated keypoint detector that eliminates the need for predefined anatomical landmarks, and this model's flexibility and accuracy in handling complex deformations make it particularly suitable for simulating and manipulating the appearances and behaviors of wasps. Our ultimate goal is to be able to reproduce a close-up video of a dyadic interaction between two wasps with the faces (and possibly the abdomens) of the original wasps replaced by those of other wasps. Toward this end, we start with a simpler computer vision problem: *given a close-up video of a moving wasp face, create a facial identity swap while preserving the original facial motion*[2].

[2]Sample videos (and source code for experiments) can be found at: https://github.com/PavlicLab/CVPR2024-CV4Animals2024-TPSM_for_Paper_Wasps/.

2. Proposed methodology

Prior models for image motion transfer commonly rely on a priori knowledge such as labels and landmarks [1, 4, 12, 21]. Such requirements are particularly difficult in insect studies, as this method necessitates potentially costly domain expertise. Furthermore, even within closely related taxonomic groups like *Polistes*, salient features for one species may be very different than those for another (e.g., Figure 1), and so highly valuable manual annotations in one study system may not easily generalize to another closely related study system. In contrast, the TPS model [22] incorporates a keypoint detector that automatically predicts pairs of keypoints to guide the mapping of motion from a source image along a driving video of desired motion. Furthermore, for enhanced flexibility and adaptability relative to methods that use linear transformations, the TPS motion model introduces a trainable non-linear methodology for modeling complex motions. In particular, the core objective of TPS motion estimation is:

$$\min \iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial y^2} \right)^2 \right) \mathrm{d}x\mathrm{d}y$$

$$\text{s.t.} \quad \mathcal{T}_{tps}(P_i^{\mathbf{S}}) = P_i^{\mathbf{D}}, \quad i = 1, 2, \ldots, N,$$

where, learned transformation $\mathcal{T}_{tps}$ optimally maps keypoints $P_i^{\mathbf{S}}$ of a source image $\mathbf{S}$ to keypoints $P_i^{\mathbf{D}}$ of a corresponding driving image (or video frame) $\mathbf{D}$. To accurately model the dynamics of both background and foreground, these TPS transformations are combined with affine transformations. This integration is facilitated through the generation of contribution maps that serve to quantify the influence of each transformation on the overall optical flow. Moreover, during the model's training phase, the implementation of a dropout mechanism is crucial in preventing the model from overfitting to specific transformations, thereby enhancing its ability to generalize across varied motions. Collectively, these features of the TPS motion model help to ensure keypoints in the source image are mapped onto those within the driving image with minimal distortion during the animation process [22]. The connection of these TPS components specialized for our wasp application are summarized in Figure 2.

3. Experiments and Ablations

Dataset: Our research utilizes a unique dataset comprising both video and photographic content of *P. fuscatus* and *P. dominula* paper wasps engaged in a broad range of solitary and social behaviors from diverse camera angles. For this study, we focus on video data of harnessed wasps that

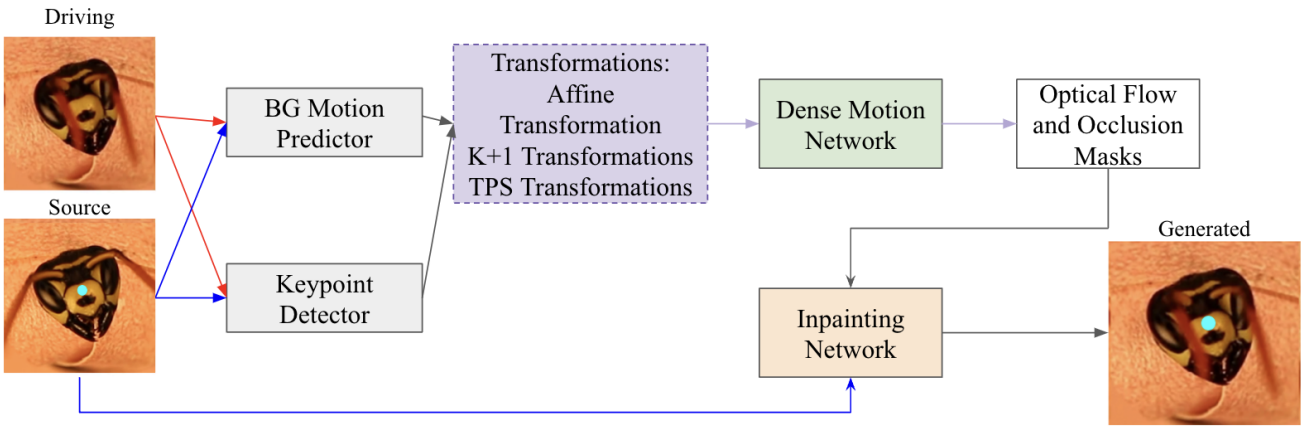

Figure 2. Summary of the TPS motion model tailored for the wasp motion-transfer application (adapted from [22]). The BG motion predictor uses affine transforms to model background motion between the source and driving images. The Keypoint Detector identifies K sets of keypoints, each leading to a TPS transformation. The Dense Motion Network merges these $K + 1$ transformations to estimate optical flow and generate multi-resolution occlusion masks. The source image is processed by the Inpainting Network, where feature maps are warped using optical flow, masked by occlusion masks.

highlights the wasps' faces, including motion of the head, mandibles, and antennae. We prepared the data for our experiments through a meticulous preprocessing stage that involved cropping each video to ensure that the wasp faces remained centrally framed throughout, which was critical for facilitating accurate face swaps. We trained TPS model on a frame-by-frame basis using 14 videos and 12,318 frames of *P. fuscatus* faces and 18 videos and 160,988 frames of *P. dominula* faces. This granularity allowed for detailed analysis and modeling, as each frame represents a potential data point for training the model to recognize and swap faces accurately.

Source Image: Ideally, the source image should be from a wasp other than the one in the driving video. However, to simplify the visual evaluation of how well a source image of a wasp face was correctly represented in the generated video (i.e., to help us identify fundamental challenges with using the TPS model in a wasp application), we used an image of the wasp face from the driving video augmented with a single blue dot on the upper quadrant of the wasp's face (Figure 2). The dataset does not contain any wasps with a blue dot, and therefore this strategy allows us to distinguish the movement in the generated output video from the driving video.

Experimental Results: Figure 3 illustrates the outcomes of training the TPS model on a dataset comprised of one video (9348 frames). Case A demonstrates the application of the pre-trained model to a *P. dominula* source image alongside a human driving video (A, D_1–D_3), and the consequent output video (A, O_1–O_3). In Case B, the pre-trained model is employed with a *P. dominula* source image and a driving video also from *P. dominula* species (B, D_1–D_3), producing the output video (B, O_1–O_3). Case C involves the model, specifically trained on *P. dominula*, processing a *P. dominula* source image and driving video (C, D_1–D_3), with the output (C, O_1–O_3). These cases provide a comparative analysis of the model's performance across different combinations of source and driving videos, highlighting its adaptability and effectiveness in generating realistic animations.

As evident in Figure 3–A, the initial experiment employed a pre-trained model using a wasp source image and a human driving video. This model was pre-trained on the VoxCeleb [9] dataset. The pre-trained TPS motion model exhibits remarkable capabilities in capturing and translating a broad spectrum of motions from a given driving video to a corresponding source image. Our initial experimentation involved employing a source image derived from the *Polistes* paper wasp dataset, upon which we applied the pre-trained model equipped with its default parameters, coupled with a standard human driving video as the motion source. The efficacy of the existing model was prominently displayed through its adeptness in mapping the complex motion dynamics from the human subject to the wasp imagery Figure 3–A. Notably, the model demonstrated a pronounced proficiency in aligning the human's forehead region with the upper head portion of the wasp in the source image, as can be seen in the accurate rotations with minimal distortion of the wasp face following similar rotations in the driving human video in Figure 3–A. This alignment showcases the model's strength in identifying and adapting to the spatial correspondences between vastly different

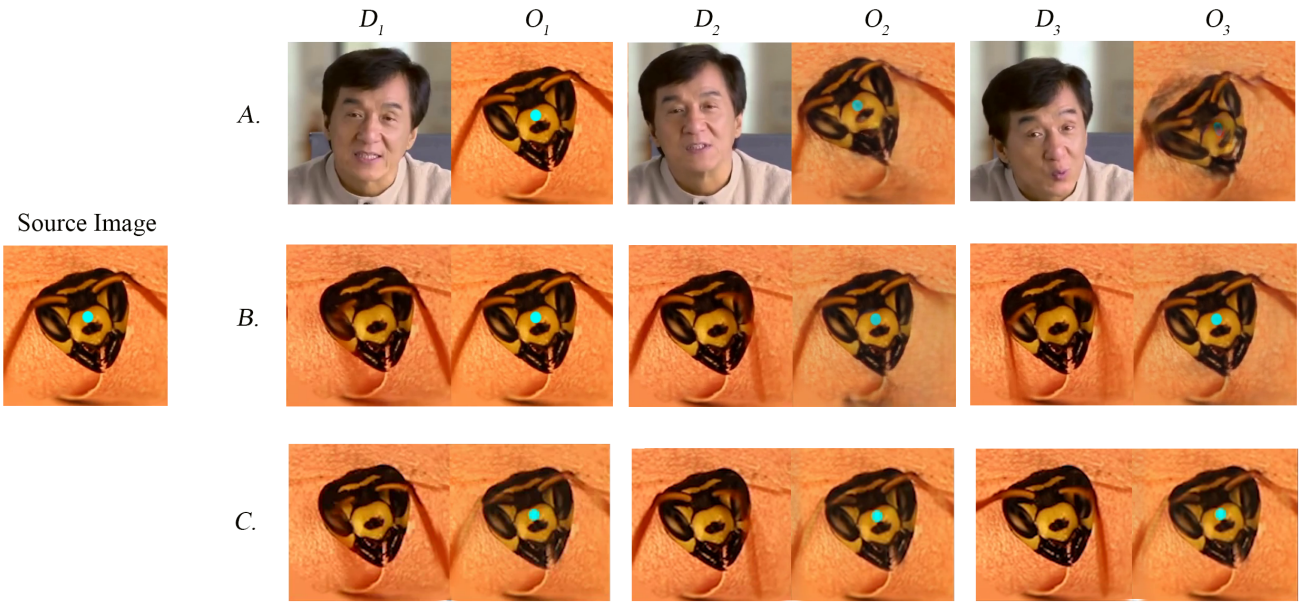

Figure 3. Qualitative comparisons of driving and output video frames. Comparison of frames from the driving video (D_x) with the corresponding output video frames (O_x) in three scenarios: Case A (top) shows results using a model pre-trained with human data and applied to a *P. dominula* source and human driving video. Case B (middle) uses the same pre-trained model but with *P. dominula* source and driving videos. Case C (bottom) applies a model custom trained on wasp faces with *P. dominula* source and driving video. Sample videos are available at: https://github.com/PavlicLab/CVPR2024-CV4Animals2024-TPSM_for_Paper_Wasps/.

anatomical structures, thereby underscoring its versatility and potential applicability in cross-species behavior studies. However, the notable differences in the bottom half of the human and wasp face (e.g., human mouth and wasp mandibles) seemed to lead to significant distortion of the wasp face in response to nuanced facial expressions or minor movements (A, D_3 & O_3).

When the pre-trained model was applied using a wasp source image paired with a wasp driving video instead, as observed in Figure 3–B, the output contained a border glitch effect. This artifact suggests a misalignment in the model's processing, potentially due to the vastly different motion patterns present in wasp behaviors compared to human facial movements. Notably, the model did not register any significant movement, indicating its difficulty in detecting and replicating the subtle and rapid motions characteristic of wasp behavior.

In contrast, the custom model trained on wasp imagery applied to a wasp driving video managed to detect mandible movement (Figure 3–C), which is a critical aspect of wasp facial expression and behavior. However, reproduction of other movements of the head may have been confounded by the motion of the antennae, which periodically move into the frame and obstruct the view of the face, leading to an inaccurate representation of intended facial swaps. This outcome suggests that the model, while capable of identifying specific areas of movement, may struggle with distinguishing between different types of isolated movements within the same video frame. The inability of the produced motion transfer to achieve a realistic portrayal of the wasp's identity swap underscores the need for further model refinement. For example, to address some of the identified issues above, future iterations of the experiment will involve cropping out the antennae from the driving and training videos to improve the model's focus on relevant facial features as well as utilizing a supercomputer to accommodate a greater number of data frames [8].

4. Discussion and Conclusion

This study explored the potential application of image-animation technology for entomological research, specifically within the context of studying the role of social-insect perception in the maintenance of dominance hierarchies. The ablation study provided insight into how various components of the TPS model contribute to the model's efficacy in simulating wasp behaviors accurately. Although the initial results did not achieve deceptive realism in animating wasp faces, they were nonetheless promising. The experiments underscored the potential for significant improve-

ments in model performance with additional training on specially prepared datasets tailored for non-human, ethological studies. Furthermore, complications that we observed that were idiosyncratic to wasps, such as the problem of faces frequently being obscured by moving antennae, may provide motivation for new kinds of motion-transfer models that not only better facilitate work with wasps but other kinds of motion transfer in human applications too.

The simulation of the likeness of a wasp's face in video marks a significant step forward toward generating representations of wasps that are recognizable to other wasps so that human experimenters can manipulate social information flow in studies of animal social networks. Future steps include expanding to include not only the face but the abdomen, which has been identified to be important to *Polistes* wasps as well [15]. Additionally, these technologies need to be applicable to driving videos with multiple interacting wasps that are freely moving. It is known that eavesdropping wasps attend to social information from behind a plastic partition [17], and there is anecdotal evidence that wasps will also attend to video information (E. Tibbetts, personal communication). Furthermore, the Asian honey bee *Apis cerana* (which is taxonomically close to the paper wasp as it is also a social hymenopteran insect) has been shown to perform a hornet-specific defense behavior to video playback of hornets on consumer-grade tablet screens [2], which shows that video playback is able to produce information that is still salient to the insect eye. However, testing the efficacy of artificially generated videos of synthetic wasp behaviors must ultimately be done to demonstrate the feasibility of this approach.

The use of virtual-reality systems is not unprecedented in studies of non-human behavior. Mice physically fixed in space can be given dynamically generated visual inputs that lead them to move as if they were free [e.g., 3, 11]. Furthermore, free-flying fruit flies can be made to fly around fictitious obstacles through proper application of dynamic images projected onto walls [5, 6], and these methods have been generalized for use with freely moving fish and mice [14]. Computer animation has also been used to produce 3D models of fish to study the role of specific morphological features in social interactions [7, 13, 20]. However, to our knowledge, this is the first example of the application of high-fidelity, generative models as a tool for studying complex social behaviors. We hope this example can motivate the creation of other new ways for Computer Vision to benefit behavioral ecology more broadly.

Ethical Considerations in Research Conduct

This study's foundational ethos revolves around the principled acquisition and utilization of data, alongside the judicious application of advanced technological tools, specifically generative motion-transfer technology. The subsequent sections delineate the ethical guidelines adhered to, ensuring the integrity and ethical conduct of the research.

Data Collection and Use: In the realm of scientific inquiry, the sanctity of data collection processes and the subsequent use of such data are paramount. This research underscores the necessity for all data collection methods—particularly those involving invertebrates *Polistes fuscatus* and *P. dominula* and other biological subjects—to be conducted ethically. Although experiments were conducted using invertebrates that are not specially regulated for scientific use, we worked to ensure minimal impact on the subjects and their natural behaviors. The data collection and husbandry methodologies employed are designed to respect the subjects' privacy and integrity, avoiding any unnecessary stress or alteration to their natural state.

Misuse Prevention: The use of generative motion-transfer technology is explicitly aligned with the research objectives, focusing on enhancing the understanding of *Polistes* paper-wasp social dynamics without deviating into unethical applications of the technology. This approach reflects a commitment to leveraging technological advancements for scientific progress while recognizing and addressing the ethical implications associated with such tools.

References

[1] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[2] Shihao Dong, Ken Tan, and James C Nieh. Visual contagion in prey defence signals can enhance honest defence. *Journal of Animal Ecology*, 90(3):594–601, 2021. 1, 5

[3] RM Douglas, NM Alam, BD Silver, TJ McGill, WW Tschetter, and GT Prusky. Independent visual threshold measurements in the two eyes of freely moving rats and mice using a virtual-reality optokinetic system. *Visual neuroscience*, 22 (5):677–684, 2005. 5

[4] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14398–14407, 2021. 2

[5] Steven N Fry, Nicola Rohrseitz, Andrew D Straw, and Michael H Dickinson. Trackfly: virtual reality for a behavioral system analysis in free-flying fruit flies. *Journal of neuroscience methods*, 171(1):110–117, 2008. 5

[6] Steven N Fry, Nicola Rohrseitz, Andrew D Straw, and Michael H Dickinson. Visual control of flight speed in *Drosophila melanogaster*. *Journal of Experimental Biology*, 212(8):1120–1130, 2009. 5

[7] Spencer J Ingley, Mohammad Rahmani Asl, Chengde Wu, Rongfeng Cui, Mahmoud Gadelhak, Wen Li, Ji Zhang, Jon

Simpson, Chelsea Hash, Trisha Butkowski, et al. anyFish 2.0: an open-source software platform to generate and share animated fish models to study behavior. *SoftwareX*, 3:13–21, 2015. 5

[8] Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, William Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobawala, Thirugnanam Jagadeesan, Praful Bhargav Basani, Torey Battelle, Rebecca Belshe, Deb McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Dhruvil Deepakbhai Shah, Sean M. Dudley, Gil Speyer, and Jason Yalim. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing*, pages 296–301, New York, NY, USA, 2023. Association for Computing Machinery. 4

[9] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a large-scale speaker identification dataset. In *Interspeech*, pages 2616–2620, 2017. 3

[10] Juanita Pardo-Sanchez and Elizabeth A. Tibbetts. Social experience drives the development of holistic face processing in paper wasps. *Animal Cognition*, 26(2):465–476, 2023. 1

[11] Brad A Radvansky and Daniel A Dombeck. An olfactory virtual reality system for mice. *Nature communications*, 9 (1):839, 2018. 5

[12] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7687–7696, 2020. 2

[13] Gil G Rosenthal and Michael J Ryan. Assortative preferences for stripes in danios. *Animal Behaviour*, 70(5):1063–1066, 2005. 5

[14] John R Stowers, Maximilian Hofbauer, Renaud Bastien, Johannes Griessner, Peter Higgins, Sarfarazhussain Farooqui, Ruth M Fischer, Karin Nowikovsky, Wulf Haubensak, Iain D Couzin, et al. Virtual reality for freely moving animals. *Nature methods*, 14(10):995–1002, 2017. 5

[15] Elizabeth A. Tibbetts. Visual signals of individual identity in the wasp *Polistes fuscatus*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1499):1423–1428, 2002. 1, 5

[16] Elizabeth A. Tibbetts and James Dale. Individual recognition: it is good to be different. *Trends in Ecology and Evolution*, 22(10):529–537, 2007. 1

[17] Elizabeth A. Tibbetts, Ellery Wong, and Sarah Bonello. Wasps use social eavesdropping to learn about individual rivals. *Current Biology*, 30(15):3007–3010.e2, 2020. 1, 5

[18] Elizabeth A. Tibbetts, Juanita Pardo-Sanchez, Julliana Ramirez-Matias, and Aurore Avarguès-Weber. Individual recognition is associated with holistic face processing in *Polistes* paper wasps in a species-specific way. *Proceedings of the Royal Society B: Biological Sciences*, 288(1943): 20203010, 2021. Publisher: Royal Society. 1, 2

[19] Elizabeth A. Tibbetts, Juanita Pardo-Sanchez, and Chloe Weise. The establishment and maintenance of dominance hierarchies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1845):20200450, 2022. Publisher: Royal Society. 1

[20] Thor Veen, Spencer J Ingley, Rongfeng Cui, Jon Simpson, Mohammad Rahmani Asl, Ji Zhang, Trisha Butkowski, Wen Li, Chelsea Hash, Jerald B Johnson, et al. anyFish: an open-source software to generate animated fish models for behavioural studies. *Evolutionary Ecology Research*, 15(3): 361–375, 2013. 5

[21] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9458–9467, 2019. 2

[22] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3657–3666, 2022. 2, 3