

---

# Supplementary Materials for “E-MoFlow: Learning Egomotion and Optical Flow from Event Data via Implicit Regularization”

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Network Architecture

2 Our network adopts a simple MLP architecture that takes spatial-temporal coordinates  $(\mathbf{x}, t)$  as input  
3 and outputs optical flow signal  $\mathbf{u} = (u, v)$ . Compared to [1–6], this coordinate-based MLP implicitly  
4 represents optical flow at spatial-temporal coordinates, essentially a velocity field, without relying on  
5 explicit discrete structures (e.g., voxel grid, event count image), enabling temporally continuous and  
6 dense flow estimation.

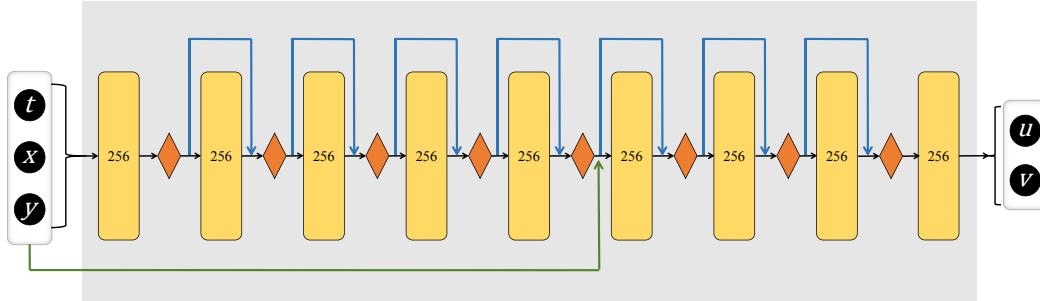


Figure 1: **Schematic Diagram of the Neural Implicit Optical Flow Field Network Architecture.** The input of the network is three-channel spatial-temporal coordinates  $(\mathbf{x}, t)$ , and the output is optical flow  $\mathbf{u} = (u, v)$ . The yellow rectangles represent 256-dimensional hidden units. The orange diamonds denote ReLU activation functions. The blue arrows indicate residual connections. The green arrows represent concatenating the original input to the output of the fifth layer.

7 Specifically, our network architecture, inspired by NeRF [7], employs 9 fully-connected layers  
8 with 256 dimensional hidden units. The first eight layers utilize ReLU activations to enforce a low  
9 Lipschitz constant, ensuring smoother responses to input variations [8], [9]. This design suppresses  
10 high-frequency features while favoring learning of low-frequency features, aligning with the prior  
11 that optical flow exhibits spatial-temporal smoothness [10, 11, 4]. Notably, no activation function  
12 (e.g., ReLU or sigmoid) is applied to the output layer, as optical flow inherently spans both positive  
13 and negative values. To further stabilize network training, we introduce residual connections between  
14 the second layer to the eighth layer and implement skip connections that concatenate the raw input  
15 with the activation outputs of fifth layer. The complete architecture is illustrated in 1.

16 Although the original NeRF architecture employs positional encoding that enhances high-frequency  
17 feature learning [7], our framework deliberately omits such encoding. This design aligns with our  
18 goal to model optical flow field which is inherently low-frequency spatial-temporal signals, while  
19 avoiding spectral bias toward high-frequency feature [12].

## 20 B Continuous Motion Representation

21 In this section, we discuss how to select an appropriate motion parameterization  $\mathcal{F}$  based on the  
 22 characteristics of camera egomotion. Given a time  $t$ ,  $\mathcal{F}$  maps it to the camera's angular velocity  $\omega$   
 23 and linear velocity  $\nu$  at that moment.

$$\mathcal{F}: t \rightarrow (\omega, \nu), \quad \mathbb{R} \rightarrow \mathbb{R}^3 \times \mathbb{R}^3 \quad (1)$$

24 In scenarios such as drones, handheld devices, and vehicle-mounted systems, camera ego-motion is  
 25 constrained by strong prior assumptions. Specifically, camera motion exhibits temporal continuity  
 26 and smoothness, meaning no abrupt changes occur within infinitesimal time intervals  $\Delta t$ . This prior  
 27 is formalized as:

$$\frac{d^k \mathcal{F}}{dt^k} \leq O_k, \quad k \in \{0, 1, 2, \dots, K\} \quad (2)$$

28  $k$  denotes the order of the derivative and  $O$  specifies the upper bounds for their respective derivatives.  
 29 The equation indicates that the  $k$ -th order motion derivatives exist and are continuous. This can be  
 30 simplified as:

$$\mathcal{F} \in \mathcal{C}^k \quad (3)$$

31  $\mathcal{C}^k$  denotes the set of functions that have continuous derivatives up to the  $k$ -th order. Additionally,  
 32 the motion of the camera is low-dimensional [13]. Thus, there is no need to over-parameterize the  
 33 camera motion (e.g., using neural networks).

34 In summary, we employ cubic B-spline as  $\mathcal{F}$  to parameterize the camera motion, as its basis functions  
 35 exhibit  $\mathcal{C}^2$  continuity and compact representation via sparse control knots [14]. Specifically, we  
 36 use four control knots  $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]^T \in \mathbb{R}^{4 \times 6}$  over a time interval  $t \in [0, 1]$ . Therefore, the  
 37 motion parameterization  $\mathcal{F}$  can be formally defined as:

$$\begin{aligned} \mathcal{F}(t) &= (\omega_\beta(t), \nu_\beta(t)) \in \mathbb{R}^3 \times \mathbb{R}^3 \\ \omega_\beta(t) &= [\mathbf{B}(t) \beta]_{0:2} \\ \nu_\beta(t) &= [\mathbf{B}(t) \beta]_{3:5} \end{aligned} \quad (4)$$

38 This definition allows us to derive the camera's angular velocity  $\omega_\beta(t)$  and linear velocity  $\nu_\beta(t)$  at  
 39 time  $t$ , where  $\mathbf{B}(t) \in \mathbb{R}^{1 \times 4}$  denotes the cubic B-spline basis functions, defined as follows:

$$\mathbf{B}(t) = \frac{1}{6} \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \quad (5)$$

40 This design choice inherently satisfies the prior assumptions: 1) Cubic B-spline intrinsically enforces  
 41  $\mathcal{C}^2$  smoothness priors, ensuring natural continuity in velocity, acceleration and jerk without requiring  
 42 explicit smoothness constraints. 2) By utilizing sparse control knots, this approach model continuous  
 43 camera motion while maintaining a low-dimensional parameterization of the 6-DoF egomotion.

## 44 C Differential Geometric Loss

45 In 3D vision, the motion of the camera  $(\omega, \nu)$  induces a motion field  $\mathbf{m}$  of projected points on the  
 46 normalized image plane  $\mathbf{x}$ . Assuming the camera is a rigid body, the relationship between the motion  
 47 field and the camera motion can be expressed by the following equation, which we formulate in  
 48 homogeneous coordinates:

$$\mathbf{m}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} A(\mathbf{x})\nu + B(\mathbf{x})\omega, \quad \mathbf{x} = [x, y, 1]^T \quad (6)$$

49 The matrices  $A(\mathbf{x})$  and  $B(\mathbf{x})$  are functions of homogeneous image coordinates defined as follows:

$$A(\mathbf{x}) = \begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \\ 0 & 0 & 0 \end{bmatrix}, \quad B(\mathbf{x}) = \begin{bmatrix} xy & -(1+x^2) & y \\ (1+y^2) & -xy & -x \\ 0 & 0 & 0 \end{bmatrix} \quad (7)$$

50 In practice,  $\mathbf{m}(\mathbf{x})$  is approximated by optical flow field  $\mathbf{u}(\mathbf{x}) = [u, v, 0]^T$  under brightness constancy  
 51 assumption.

$$\mathbf{u}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} A(\mathbf{x})\boldsymbol{\nu} + B(\mathbf{x})\boldsymbol{\omega} \quad (8)$$

52 Eq.(8) is a critically important motion field equation, which establishes the relationship between  
 53 optical flow and camera egomotion [15], [16].

54 However, the presence of  $Z(\mathbf{x})$  in this equation implies that recovering camera motion from optical  
 55 flow or deriving optical flow from camera motion requires knowledge of depth values at each image  
 56 coordinate. Prior works such as [17–21], rely on depth priors or assume locally shared depth values  
 57 when performing 6-DOF motion estimation, while methods like [4, 22, 23] jointly estimate depth  
 58 alongside optical flow and 6-DOF motion. However, this expands the parameterization space of  
 59 the optimization problem, introducing additional degrees of freedom that may lead to convergence  
 60 to local minima. Therefore, to enable the formulation of an unsupervised loss function that can  
 61 simultaneously estimate optical flow and 6-DOF motion with high accuracy, we need to eliminate the  
 62 dependence on  $Z(\mathbf{x})$ .

63 We transpose Eq.(8) and then left-multiply by  $\boldsymbol{\nu} \times \mathbf{x}$ , form the inner product of  $\mathbf{u}(\mathbf{x})$  and  $\boldsymbol{\nu} \times \mathbf{x}$ ,  
 64 yielding a scalar equation to isolate  $Z(\mathbf{x})$  as follows.  $\times$  denotes the cross product operation.

$$\mathbf{u}(\mathbf{x})^T (\boldsymbol{\nu} \times \mathbf{x}) = \left( \frac{1}{Z(\mathbf{x})} A(\mathbf{x})\boldsymbol{\nu} + B(\mathbf{x})\boldsymbol{\omega} \right) (\boldsymbol{\nu} \times \mathbf{x}) \quad (9)$$

65 Simplify the above equation to obtain:

$$\mathbf{u}(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} = \frac{1}{Z(\mathbf{x})} \boldsymbol{\nu}^T A(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} + \boldsymbol{\omega}^T B(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} \quad (10)$$

66 where  $[\cdot]_{\times}$  denotes the skew-symmetric operation. Interestingly, it can be proven that the coefficient  
 67 of the term that involves  $Z(\mathbf{x})$  in Eq.(10) is identically zero.

$$\boldsymbol{\nu}^T A(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} \equiv 0 \quad (11)$$

68 Therefore, Eq.(10) can be further simplified as follows:

$$\mathbf{u}(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} - \boldsymbol{\omega}^T B(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} = 0 \quad (12)$$

69 By expanding  $\boldsymbol{\omega}^T B(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x}$ , Eq.(12) can be rewritten in the following form:

$$\mathbf{u}(\mathbf{x})^T [\boldsymbol{\nu}]_{\times} \mathbf{x} - \mathbf{x}^T \mathbf{s} \mathbf{x} = 0, \quad \mathbf{s} = \frac{1}{2} ([\boldsymbol{\omega}]_{\times} [\boldsymbol{\nu}]_{\times} + [\boldsymbol{\nu}]_{\times} [\boldsymbol{\omega}]_{\times}) \quad (13)$$

70 Finally, we obtained an equation that connects the optical flow field and camera motion without  
 71 relying on depth values. Eq.(13) can theoretically be regarded as a differential form of the epipolar  
 72 constraint. We use this as our differential geometric loss to jointly learn optical flow and 6-DoF  
 73 motion, as shown in the following equation.

$$L_{\text{geometry}}(t, \mathbf{x}, \theta, \beta) = \|\mathbf{u}_{\theta}(t, \mathbf{x})^T [\boldsymbol{\nu}_{\beta}(t)]_{\times} \mathbf{x} - \mathbf{x}^T \mathbf{s}_{\beta}(t) \mathbf{x}\|_2^2, \quad (14)$$

$$\mathbf{s}_{\beta}(t) = \frac{1}{2} ([\boldsymbol{\omega}_{\beta}(t)]_{\times} [\boldsymbol{\nu}_{\beta}(t)]_{\times} + [\boldsymbol{\nu}_{\beta}(t)]_{\times} [\boldsymbol{\omega}_{\beta}(t)]_{\times})$$

74 Here,  $\mathbf{u}_{\theta}(t, \mathbf{x})$  represents the optical flow obtained from our neural implicit representation, while  
 75  $\boldsymbol{\omega}_{\beta}(t)$  and  $\boldsymbol{\nu}_{\beta}(t)$  denote the angular velocity and linear velocity of the camera, derived from the cubic  
 76 B-spline continuous motion representation.

## 77 D More Qualitative Results

78 We further provide additional qualitative results. As shown in 2, our method achieves comprehensive  
 79 6-DoF motion estimation on the MVSEC dataset [24]. The angular velocity and linear velocity  
 80 estimated by our approach closely match the ground-truth motion.

81 We provide additional qualitative comparisons of optical flow estimation between our method and  
 82 MultiCM [10], the second-best performing baseline. As shown in 4, our approach predicts optical  
 83 flow with superior continuity and smoothness, validating the effectiveness of our neural implicit  
 84 optical flow field representation. The color wheel used to visualize optical flow is shown in 3, where  
 85 different colors encode the magnitude and direction of the optical flow.

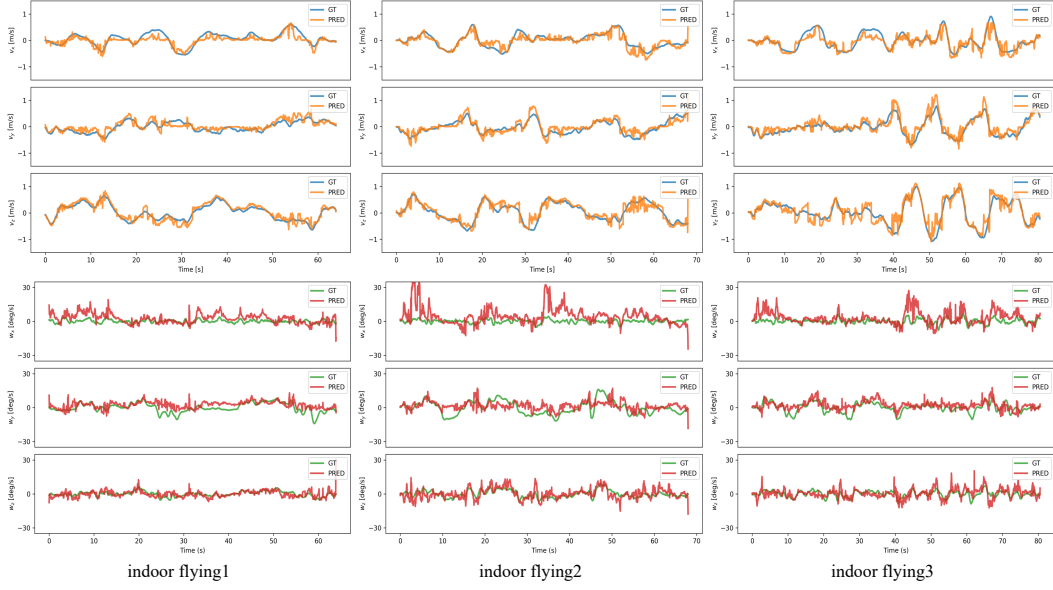


Figure 2: **Complete qualitative results of 6-DoF motion estimation on the MVSEC dataset.** The top section displays the linear velocity estimation results (in  $m/s$ ), while the bottom section shows the angular velocity estimation results (in  $deg/s$ ). The top, middle, and bottom rows in each subfigure correspond to the  $x$ -axis,  $y$ -axis, and  $z$ -axis results, respectively.



Figure 3: **Color wheel for visualizing optical flow.** A green color in the optical flow visualization corresponds to motion directed toward the lower-left corner of the image, while the saturation of the color encodes the flow magnitude — more vivid hues indicate larger displacement values.

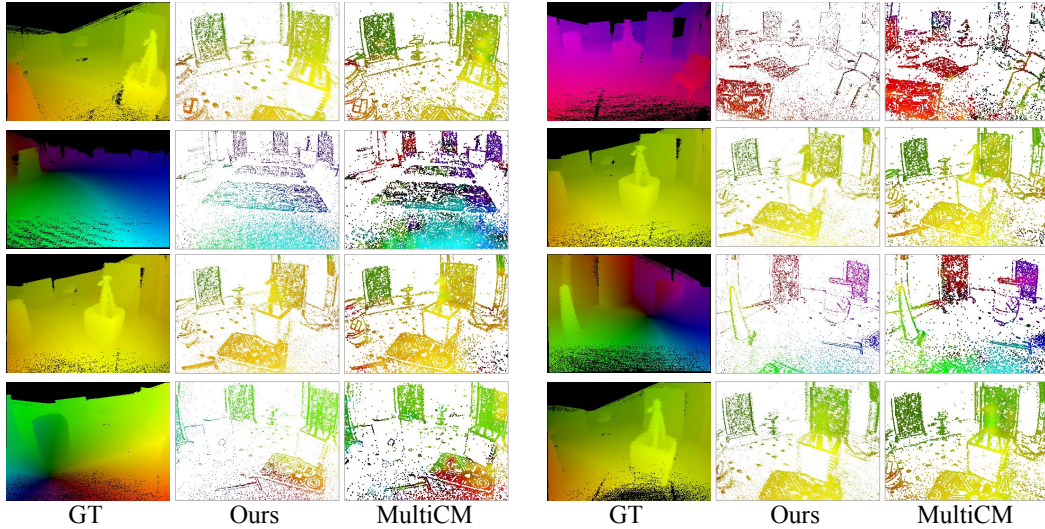


Figure 4: **More qualitative results of optical flow estimation on the MVSEC dataset.** It can be clearly observed that our method estimates smoother optical flow, free from abrupt variations, and demonstrates closer alignment with the ground truth optical flow. This indicates that our approach more effectively models the intrinsically spatial-temporally continuous optical flow field.

## References

- [1] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *European Conference on Computer Vision*, pages 18–37. Springer, 2024.
- [2] Hao Zhuang, Zheng Fang, Xinjie Huang, Kuanxu Hou, Delei Kong, and Chenming Hu. Ev-mgrflownet: Motion-guided recurrent network for unsupervised event-based optical flow with hybrid motion-compensation loss. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.
- [3] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *Advances in Neural Information Processing Systems*, 34:7167–7179, 2021.
- [4] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019.
- [5] Federico Paredes-Vallés, Kirk YW Scheper, Christophe De Wagter, and Guido CHE De Croon. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9695–9705, 2023.
- [6] Federico Paredes-Vallés and Guido CHE De Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021.
- [7] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [8] Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. *Advances in Neural Information Processing Systems*, 33:7344–7353, 2020.
- [9] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.
- [10] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022.
- [11] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.
- [12] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019.
- [13] Richard Hartley. *Multiple view geometry in computer vision*, volume 665. Cambridge university press, 2003.
- [14] Larry Schumaker. *Spline functions: basic theory*. Cambridge university press, 2007.
- [15] Hugh Christopher Longuet-Higgins and Kvetoslav Prazdny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 208(1173):385–397, 1980.
- [16] Marco Zucchelli. *Optical flow based structure from motion*. PhD thesis, Numerisk analys och datalogi, 2002.
- [17] Zhongyang Ren, Bangyan Liao, Delei Kong, Jinghang Li, Peidong Liu, Laurent Kneip, Guillermo Gallego, and Yi Zhou. Motion and structure from event-based normal flow. In *European Conference on Computer Vision*, pages 108–125. Springer, 2024.
- [18] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018.
- [19] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019.

- 136 [20] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Event collapse in contrast maximization  
137 frameworks. *Sensors*, 22(14):5190, 2022.
- 138 [21] Urbano Miguel Nunes and Yiannis Demiris. Entropy minimisation framework for event-based vision  
139 model estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August*  
140 *23–28, 2020, Proceedings, Part V 16*, pages 161–176. Springer, 2020.
- 141 [22] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical  
142 flow, depth and ego-motion estimation by contrast maximization. *IEEE Transactions on Pattern Analysis*  
143 *and Machine Intelligence*, 2024.
- 144 [23] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A Yorke, and Yiannis Aloimonos. Unsuper-  
145 vised learning of dense optical flow, depth and egomotion with event-based sensors. In *2020 IEEE/RSJ*  
146 *International Conference on Intelligent Robots and Systems (IROS)*, pages 5831–5838. IEEE, 2020.
- 147 [24] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis.  
148 The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics*  
149 *and Automation Letters*, 3(3):2032–2039, 2018.