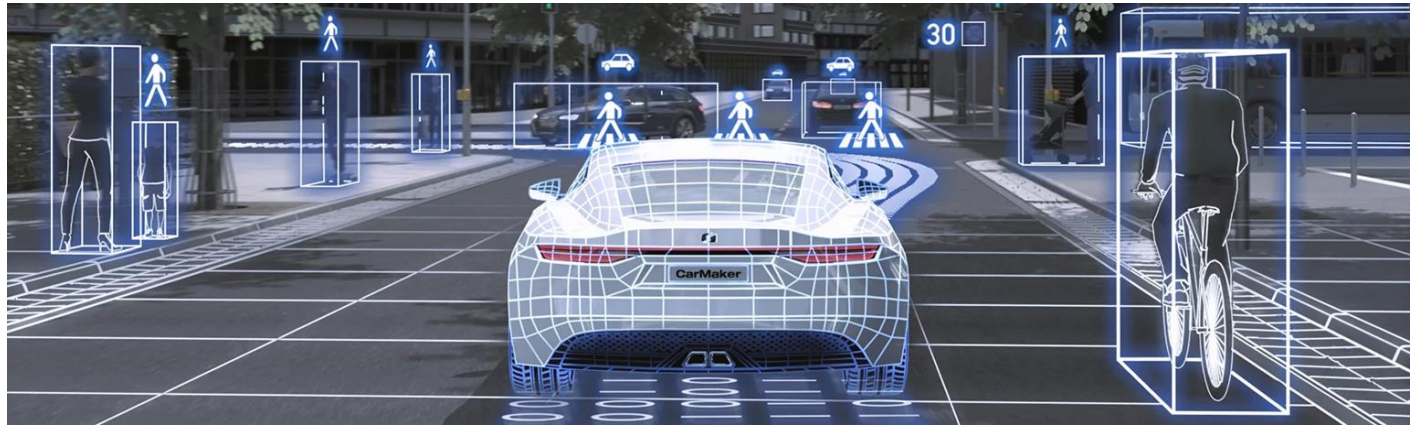
ETH zürich

EvenNICER-SLAM: Event-based Neural Implicit Encoding SLAM

Shi Chen
M.Sc. in Electrical Engineering and Information Technology
Semester Project
Supervisor: Dr. Danda Pani Paudel, Prof. Luc Van Gool
28.02.2023

Background: Dense Visual SLAM

- SLAM: Simultaneous Localization and Mapping
- Categorization: Sparse (landmarks, point cloud…) / **dense** (reconstructing every pixel/voxel)
- Input: **RGB**, **depth**, inertial, **event**…
- Applications: Autonomous driving, AR/VR, robot navigation…



(https://ipg-automotive.com/en/applications/autonomous-vehicles/)

(https://www.xmreality.com/blog/hololens2)

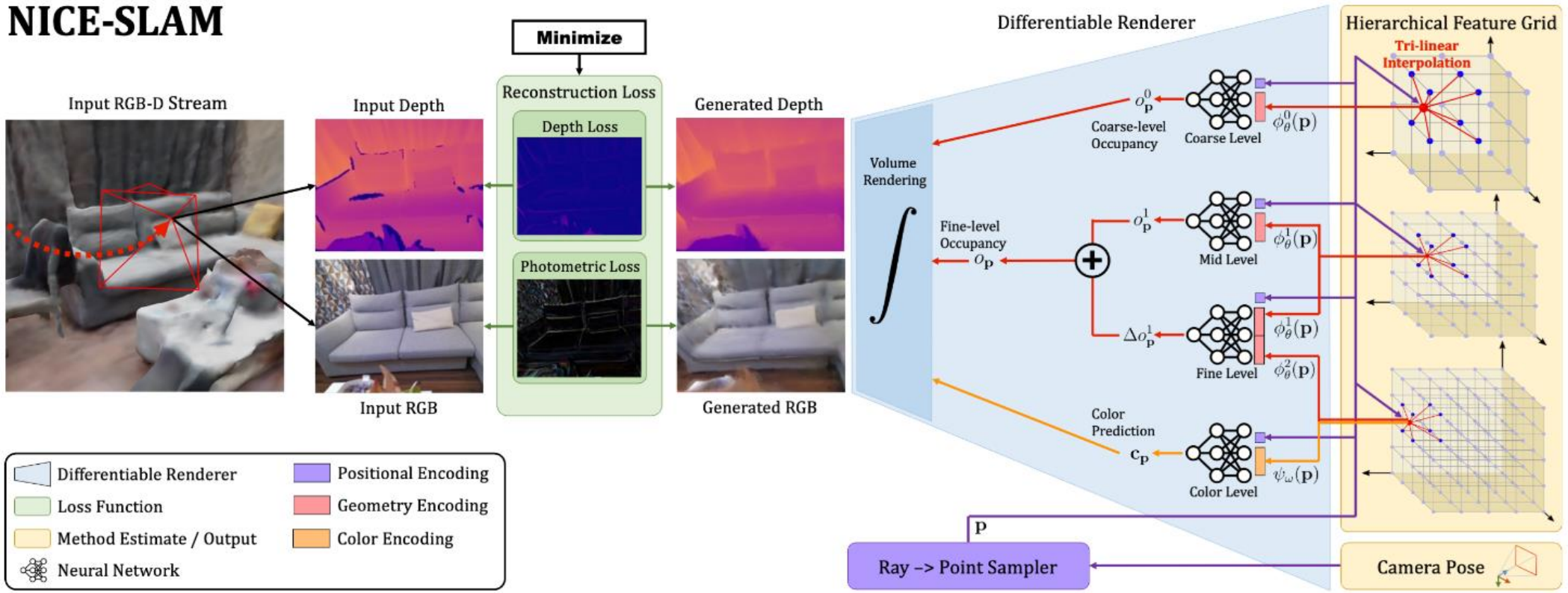
Neural Implicit Encoding SLAM: NICE-SLAM [Zhu,Peng2021]

- Dense RGB-D Neural Implicit Encoding SLAM
- Hierarchical feature grid + multi-level decoders
- Scalable to larger indoor scenes
- **EvenNICER-SLAM is based on NICE-SLAM**

(https://pengsongyou.github.io/nice-slam)

NICE-SLAM: Basis of EvenNICER-SLAM

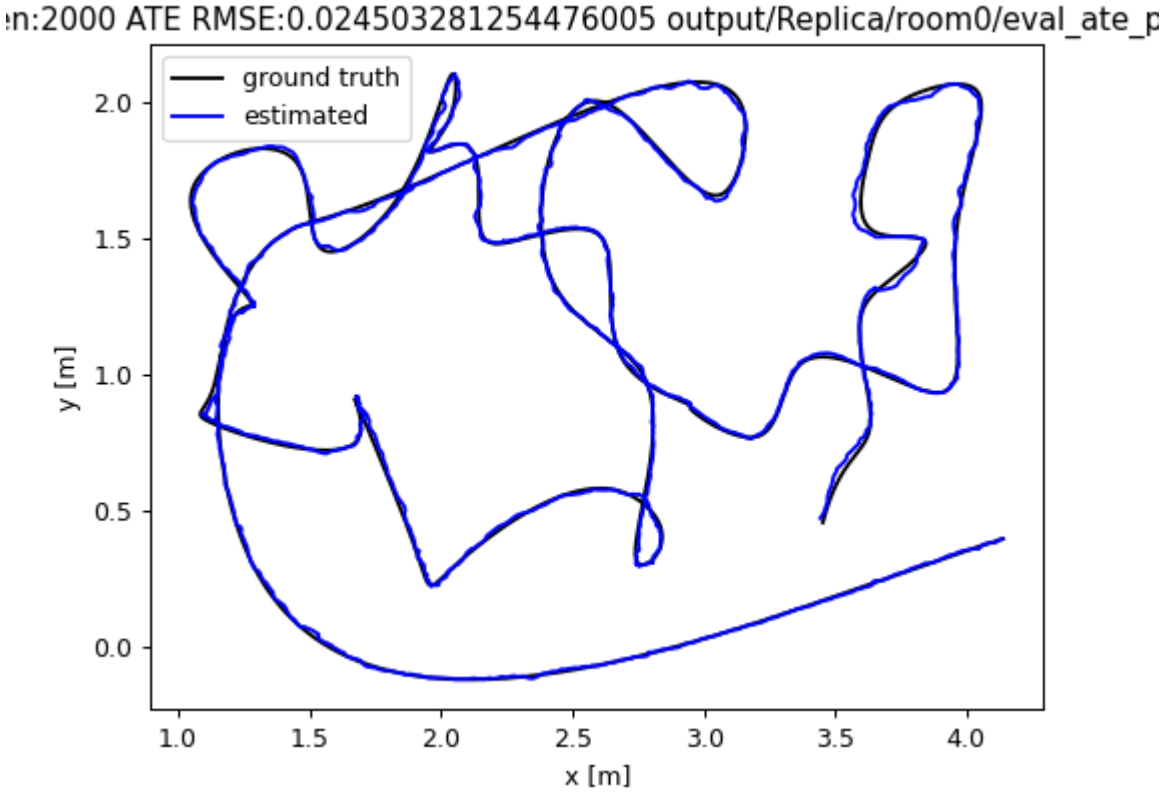
- Coarse to fine
- Occupancy & color predictions from decoders processed through volume renderer to generate depth maps & RGB images
- Depth loss & photometric loss are computed and backpropagated to optimize
  - Camera pose
  - Hierarchical Feature Grid

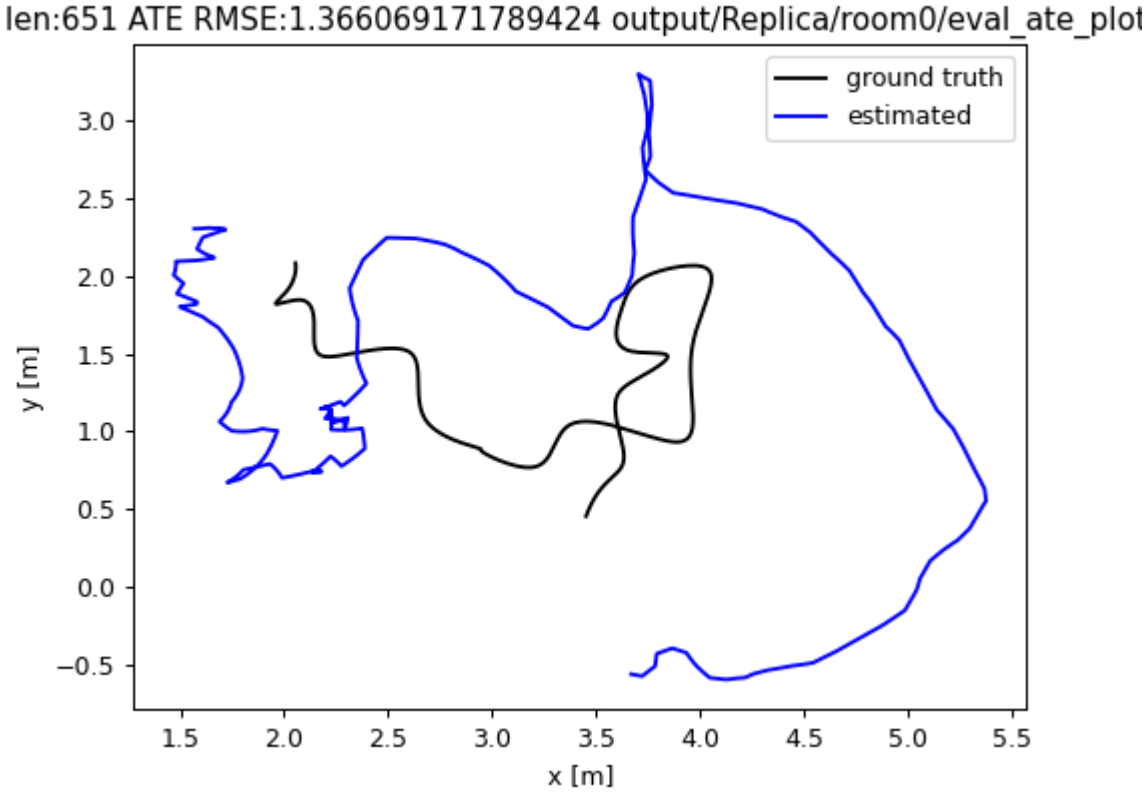


(https://pengsongyou.github.io/nice-slam)

NICE-SLAM is not “nice” enough with fast-moving cameras!

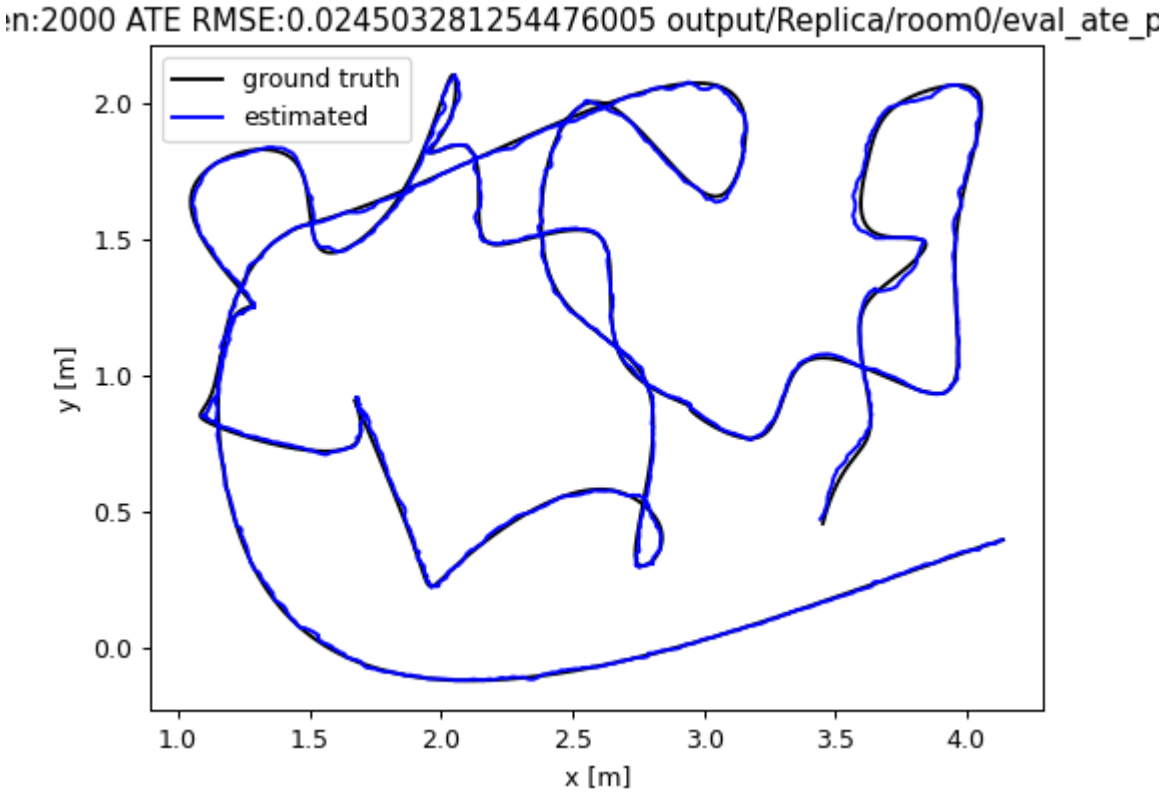
- Example: Replica dataset [Straub2019]
- 2000 RGB-D frames fully fed vs. Same sequence but only fed every fifth frame
- **NICE-SLAM shows poor tracking performance with larger inter-frame camera translation & rotation!**

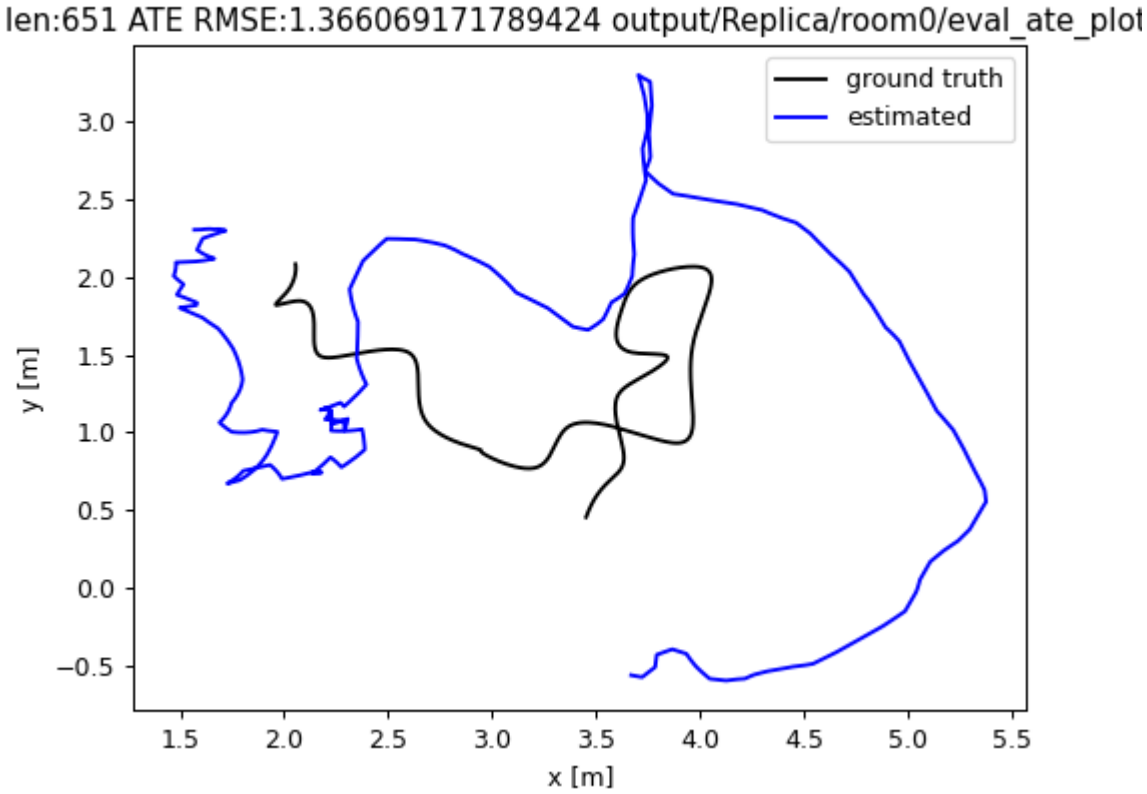




NICE-SLAM is not “nice” enough with fast-moving cameras!

- Total distance ≈ 20m
  - → average translation = total distance / n_frames ≈ 20m / 2000 frames = 1cm/frame
  - → With reduced input, average translation ≈ 5cm/frame
- **NICE-SLAM would likely crash with an average inter-frame camera pose translation of 5cm!**





Scenarios involving high-speed camera motion

- Large translation: autonomous drone racing
  - Top speed up to 30m/s [Hanover2023]
  - Most RGB-D cameras operate at 30 FPS
  - → Maximum translation ≈ 1m/frame >> 5cm/frame
  - **NICE-SLAM not going to work in this scenario!**

(https://www.youtube.com/watch?v=bR4Gq9qfpnM&ab_channel=ABCNews%28Australia%29)

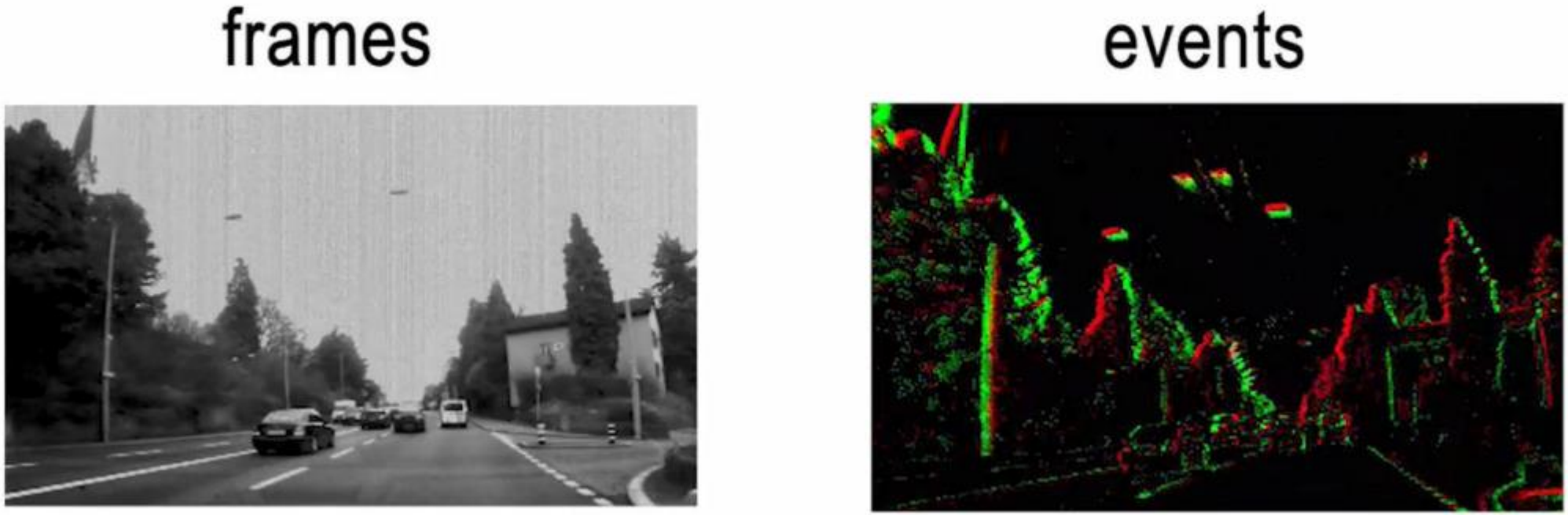
Scenarios involving high-speed camera motion

- Quick rotation: MR headsets, Microsoft Hololens 2 as an example
  - Peak speed of human head movement ≈ 240°/s [Guan2022]
  - High-resolution depth camera operates at 5 FPS
  - → Maximum rotation ≈ 48°/frame
  - **NICE-SLAM not going to work, either!**

(https://mixed.de/hololens-2-update-bringt-viele-neuerungen/)

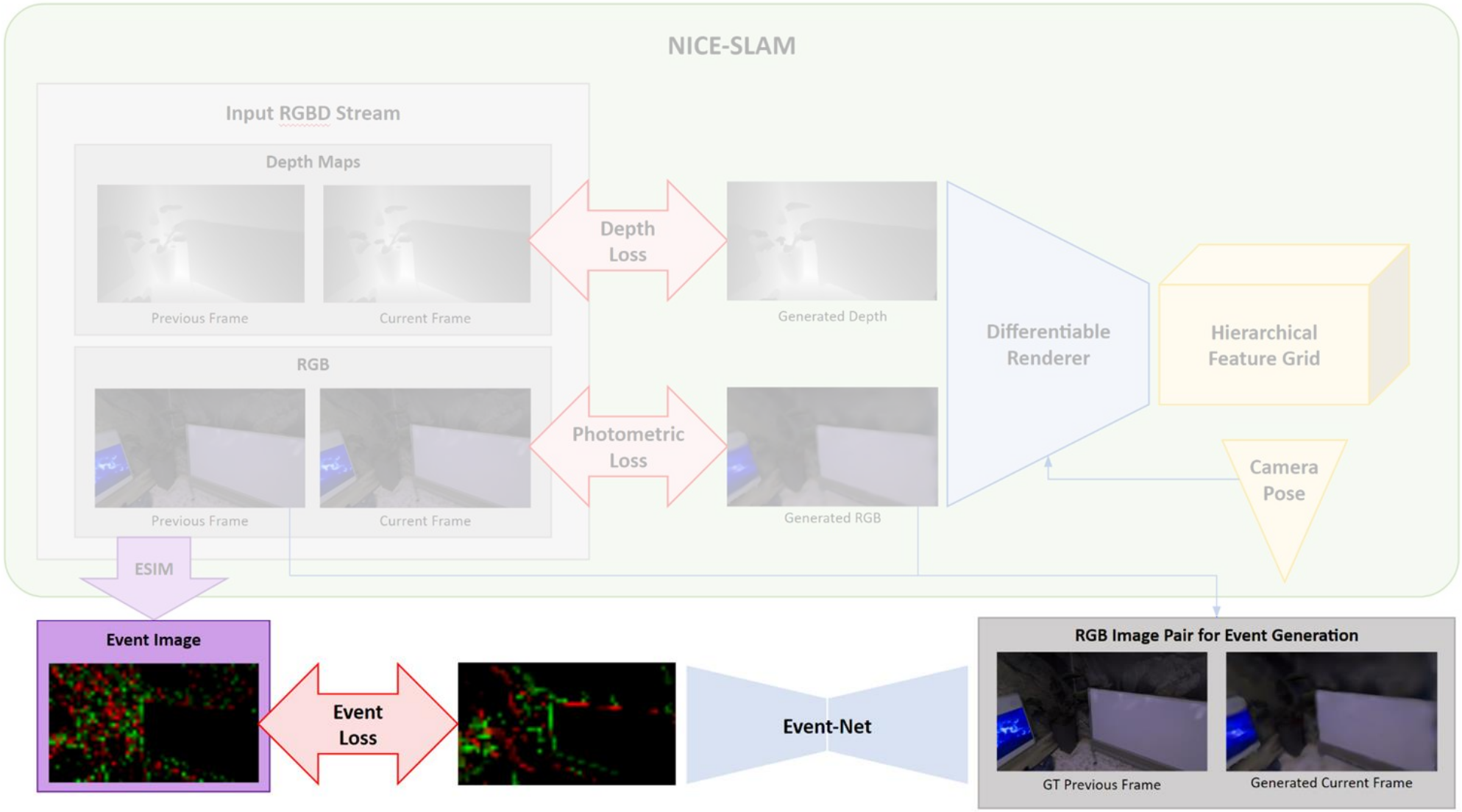
EvenNICER-SLAM: Events make NICE-SLAM “even nicer”!

- Event camera: Each pixel responds to CHANGE in intensity
  - An event is triggered when the change in log-intensity reaches "contrast threshold"
- Particularly suitable for high-speed applications
- Can we incorporate event input into NICE-SLAM to improve its robustness against high-speed camera motion?



(https://rpg.ifi.uzh.ch/docs/scaramuzza/Tutorial_on_Event_Cameras_Scaramuzza.pdf)

Overview of EvenNICER-SLAM

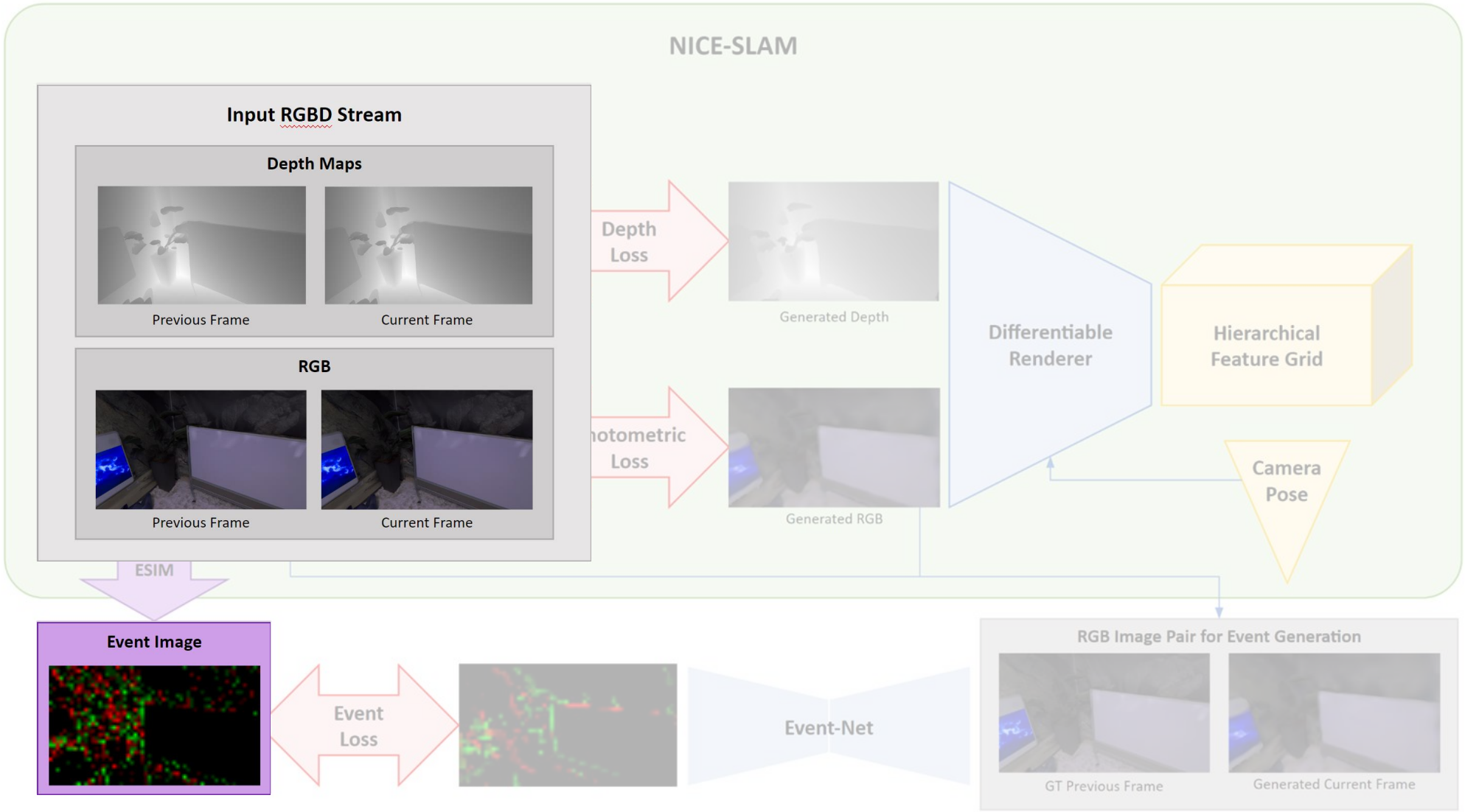


Event loss backpropagation stream

Overview of EvenNICER-SLAM

Lower frequency

Higher frequency

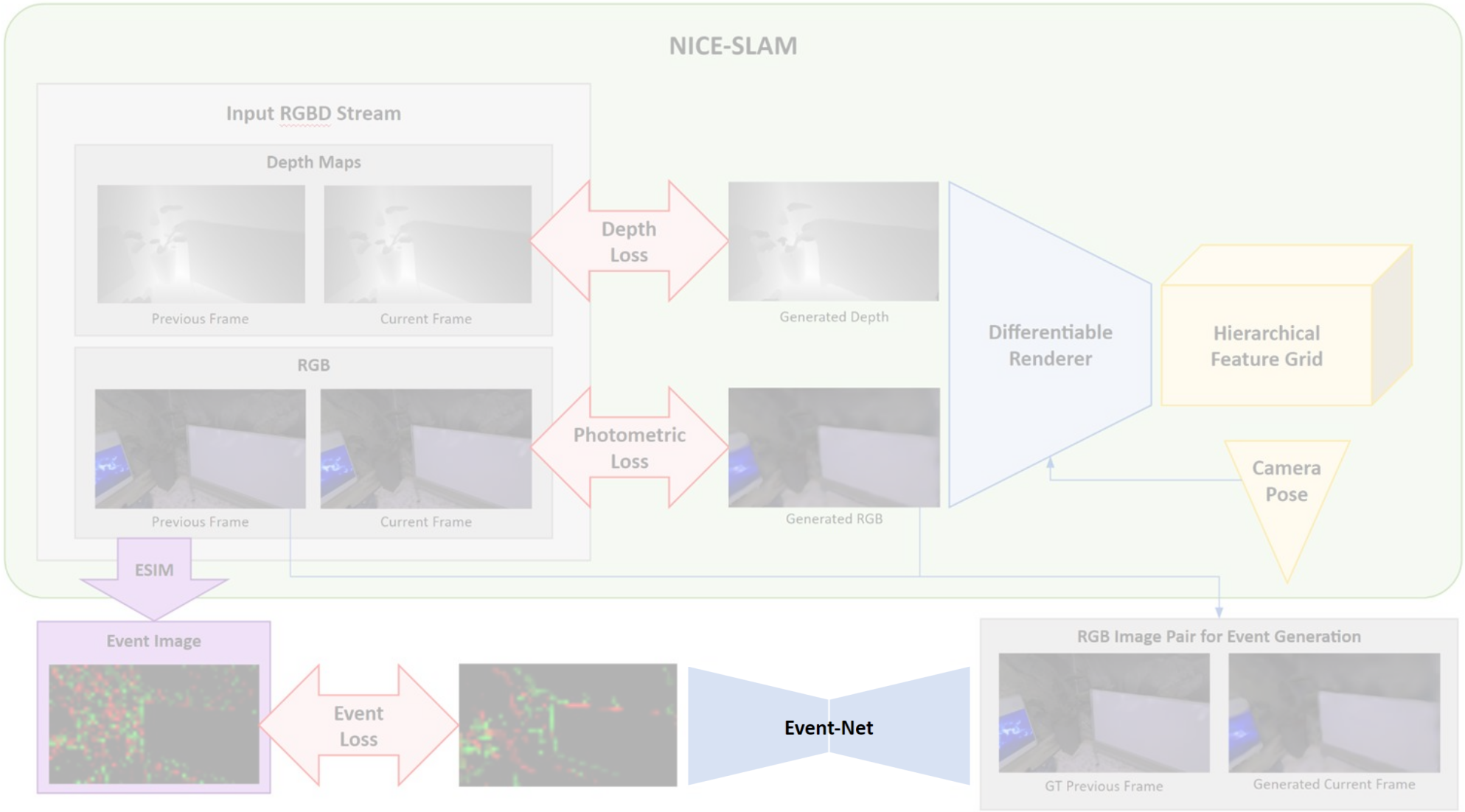


ESIM: Event Simulator [Rebecq2019]

ESIM: Event Simulator [Rebecq2019]

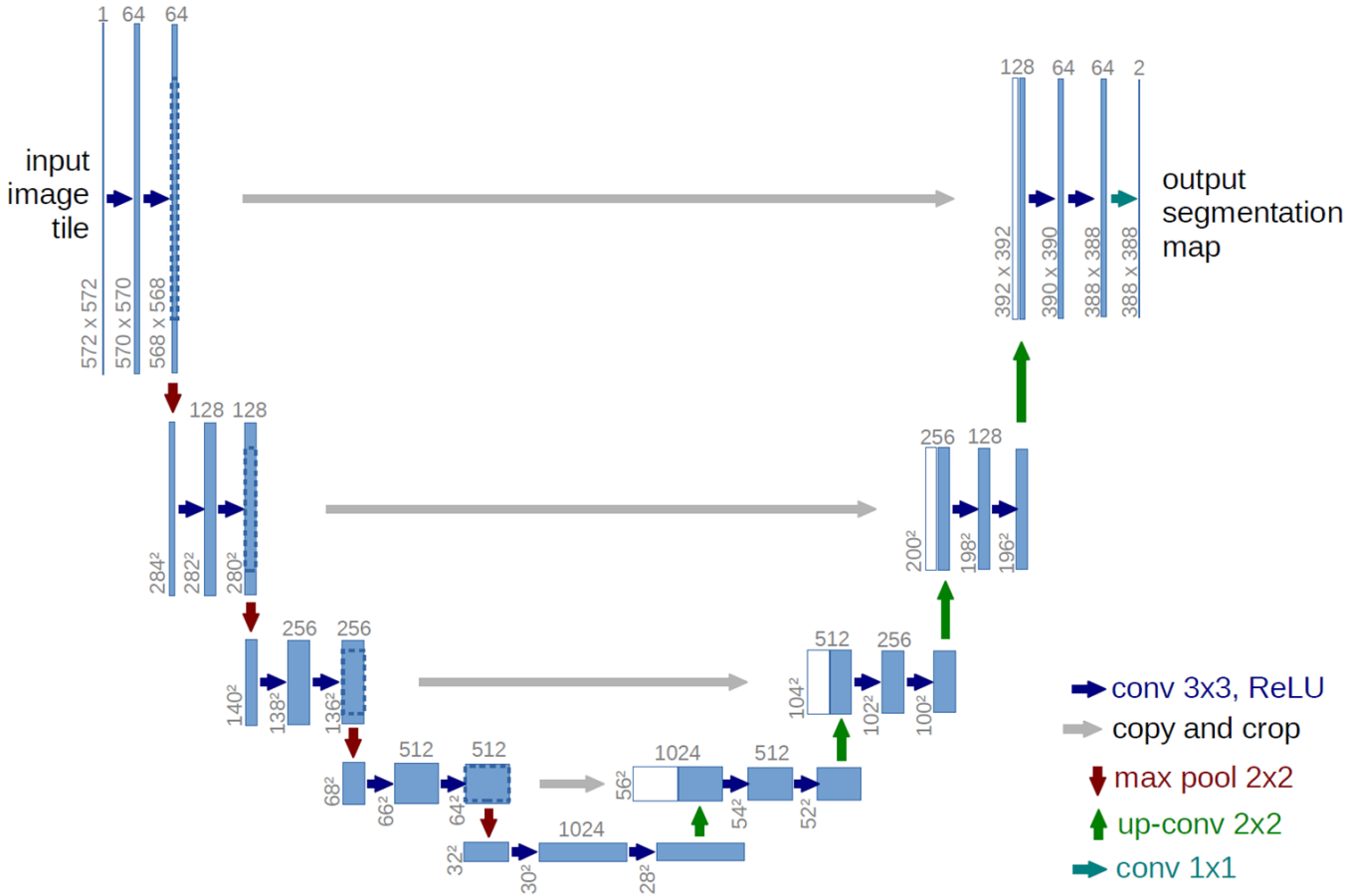
- Takes two RGB images as input and simulate events in between
- “Event image”: Integration of all events in a time interval
- Synthesize event images between all adjacent frames → GT event images

Event-Net: Differentiable Event Generator



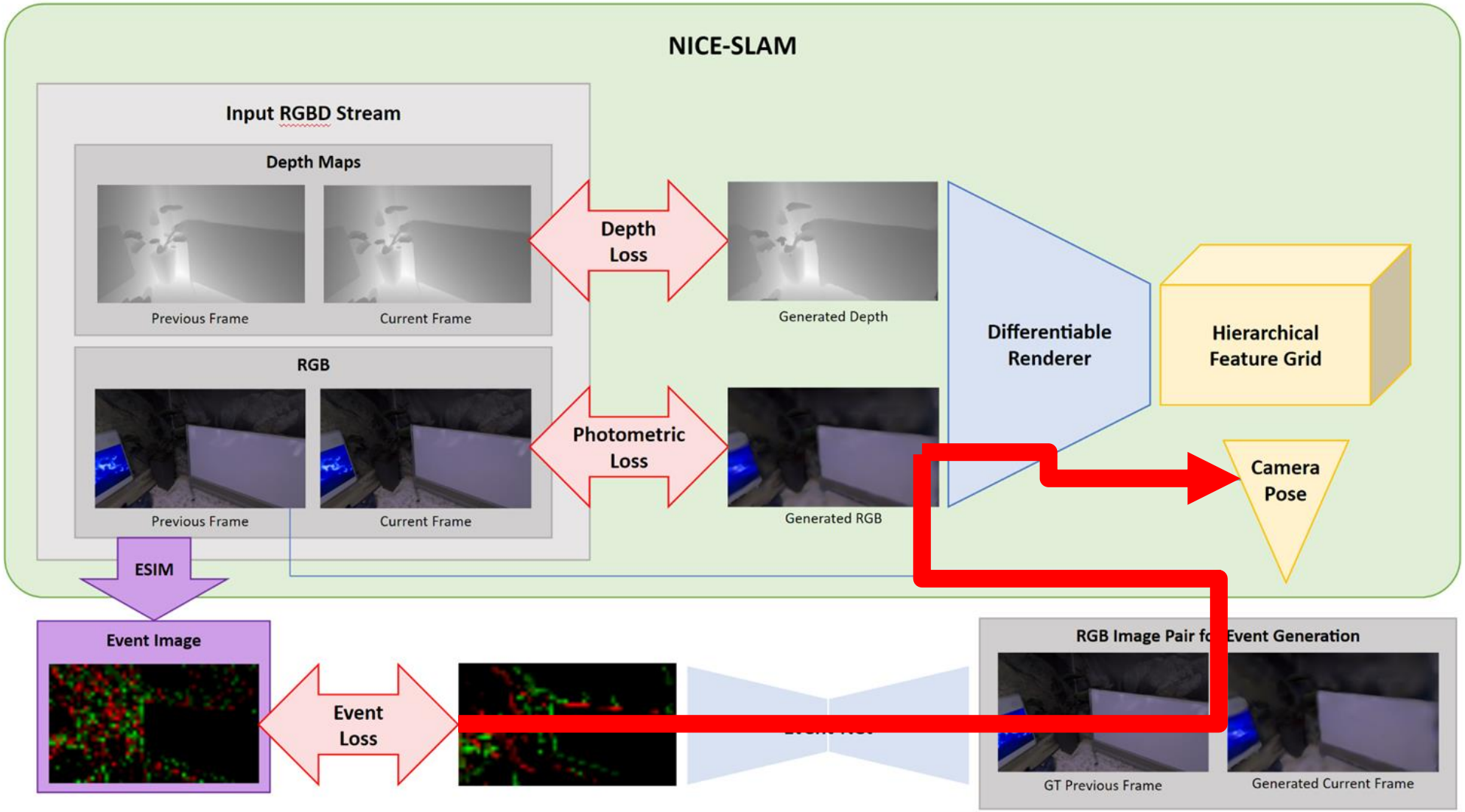
Event-Net: Differentiable Event Generator

- Differentiable so as to backpropagate “event loss”
- Architecture: U-Net [Ronneberger2015] (capable of handling high resolution features)
- We pretrain Event-Net with the GT event images



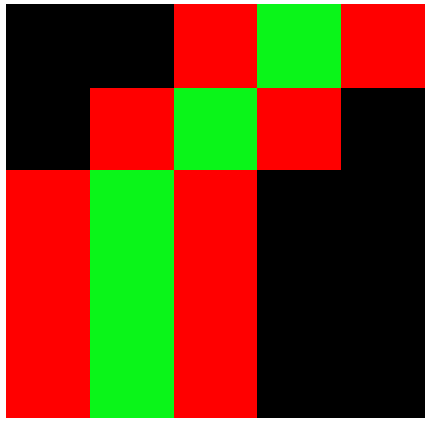
(https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/)

Optimization Using Event Loss

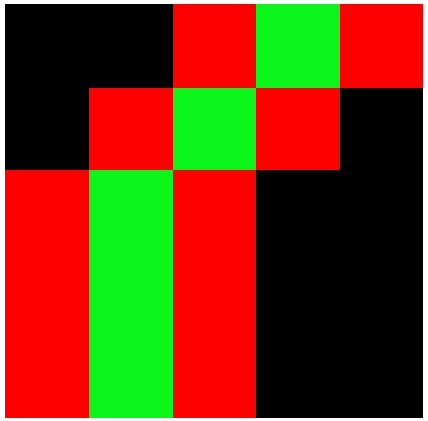


Pitfalls of Event Loss: Alignment issue

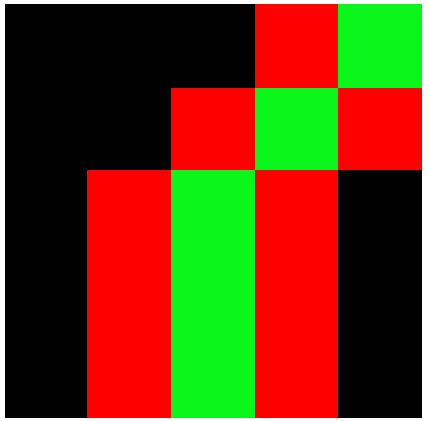
- Event loss: aims to measure the discrepancy between predicted and GT event images
- Pixelwise error?
  - Intuition: A < B < C
  - Actually: A < **C < B**
  - Rough parameter space
- → **Gaussian filter** to both predicted & GT event images



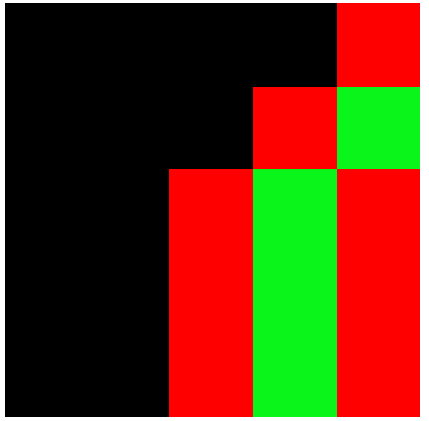
GT event image



Prediction A
(perfectly aligned)



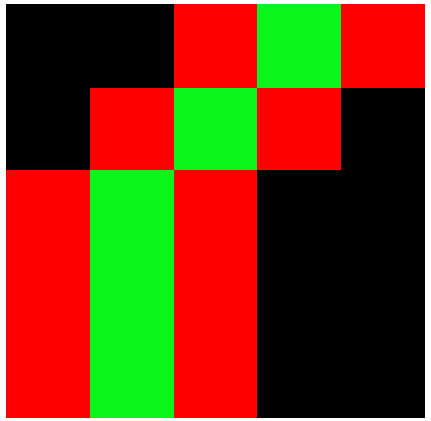
Prediction B
(shifted by 1 pixel)



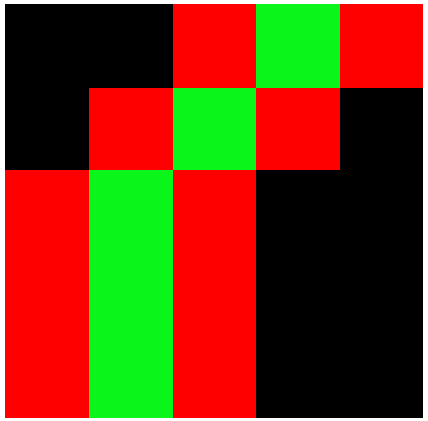
Prediction C
(shifted by 2 pixel)

Pitfalls of Event Loss: Alignment issue

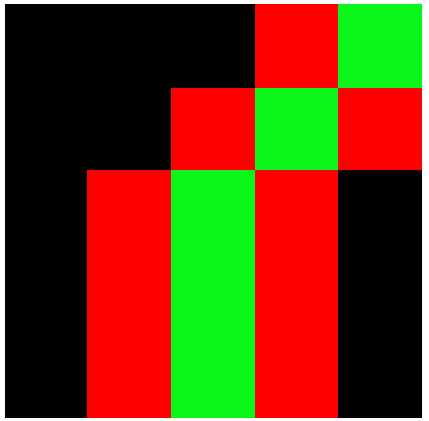
- Event loss: aims to measure the discrepancy between predicted and GT event images
- Pixelwise error?
  - Intuition: A < B < C
  - Actually: A < **C < B**
  - Rough parameter space
- → **Gaussian filter** to both predicted & GT event images



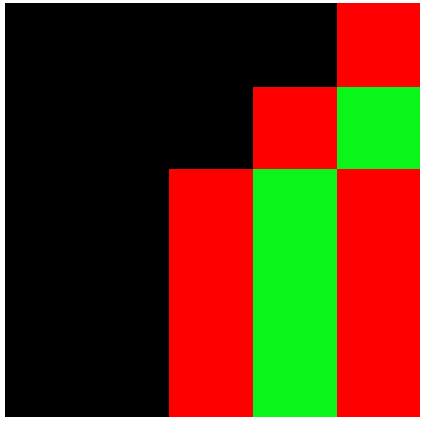
GT event image



Prediction A
(perfectly aligned)



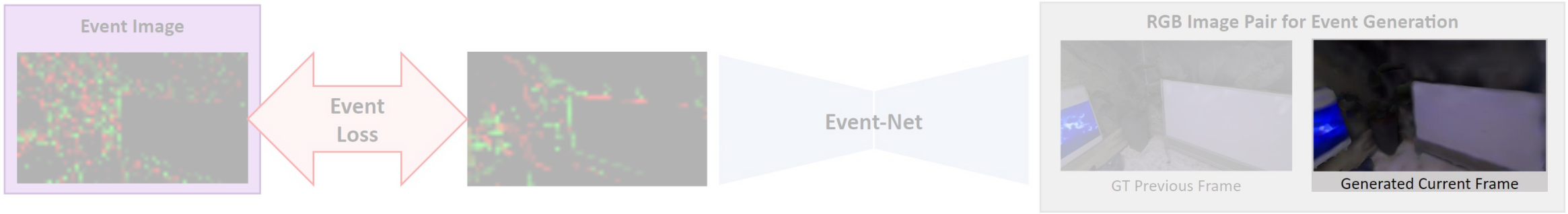
Prediction B
(shifted by 1 pixel)



Prediction C
(shifted by 2 pixel)

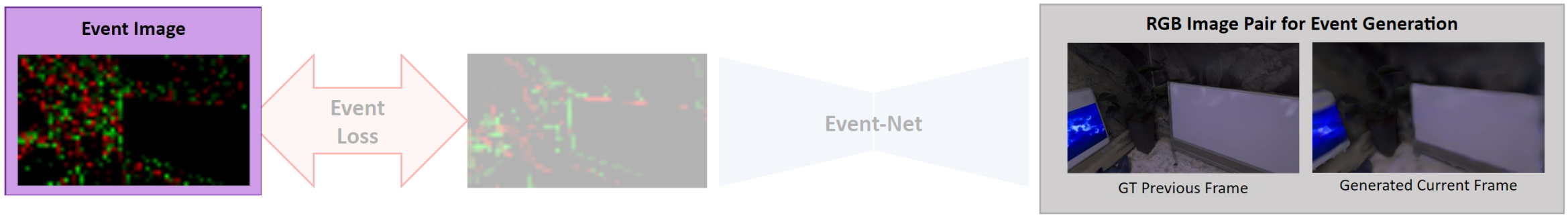
Pitfalls of Event Loss

- Pitfall #2: Unaffordable computational cost of rendering
  - → Only render a largely downsampled image, downscale GT event images accordingly



Pitfalls of Event Loss

- Pitfall #2: Unaffordable computational cost of rendering
  - → Only render a largely downsampled image, downscale GT event images accordingly
- Pitfall #3: Error accumulation
  - → Always use the latest GT RGB image as the first input of Event-Net
  - Simply sum up corresponding GT event images



Definition of Event Loss

- We have predicted event images:

Event-Net mapping

$$\widehat{I}^{t}_{event} = f(I^{t_{GT}}_{RGB}, \widehat{I}^{t}_{RGB}),\ t - \tau \leq t_{GT} \leq t - 1$$

Timestamp of most recent GT RGB-D

- And L2 event loss:

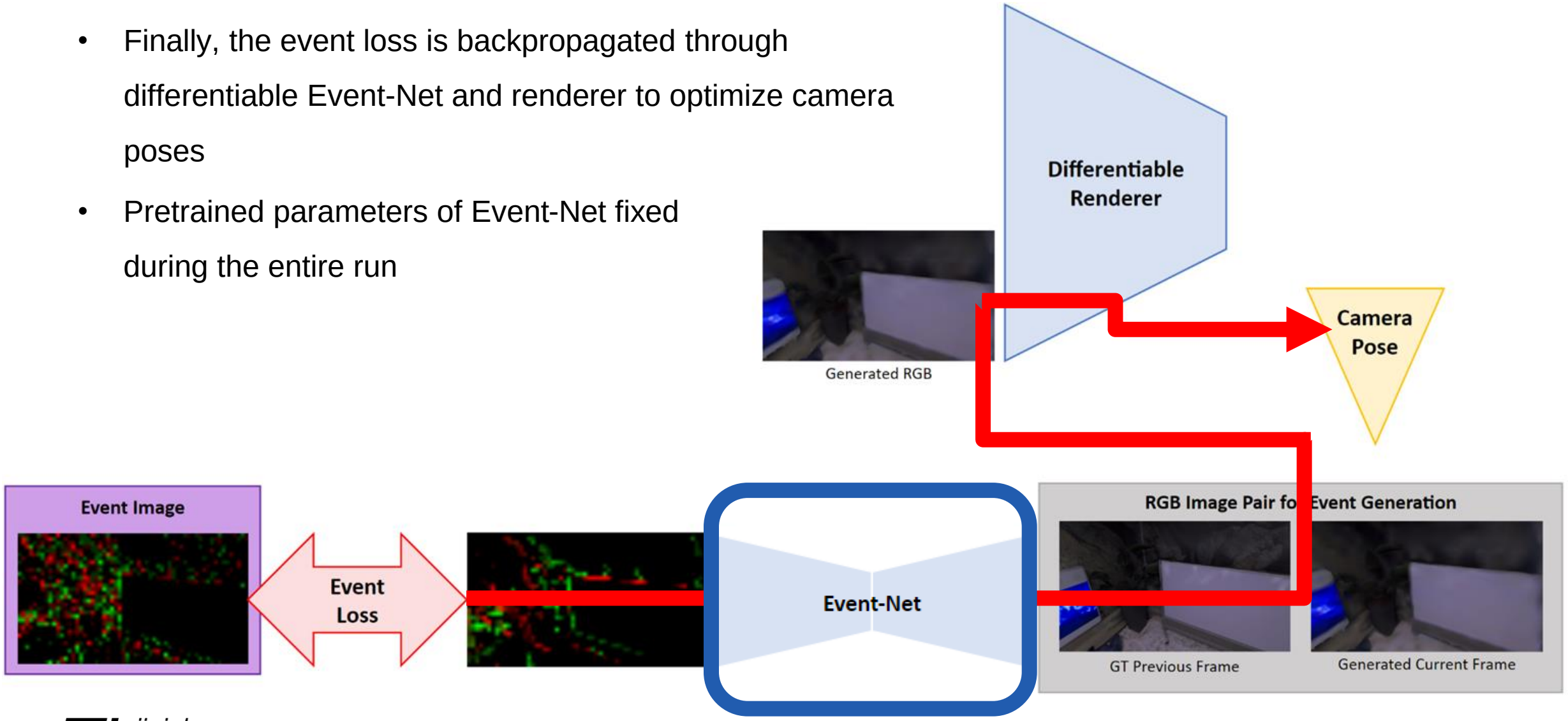
Balancing coefficient

Sum of GT event images in between

$$L_{event} = \lambda_{event} \sum_{x} \sum_{y} \left(Gaussian \left(\sum_{t=t_{GT}+1}^{t} I^{t}_{event}(x,y) \right) - Gaussian \left(\widehat{I}^{t}_{event}(x,y) \right) \right)^2$$

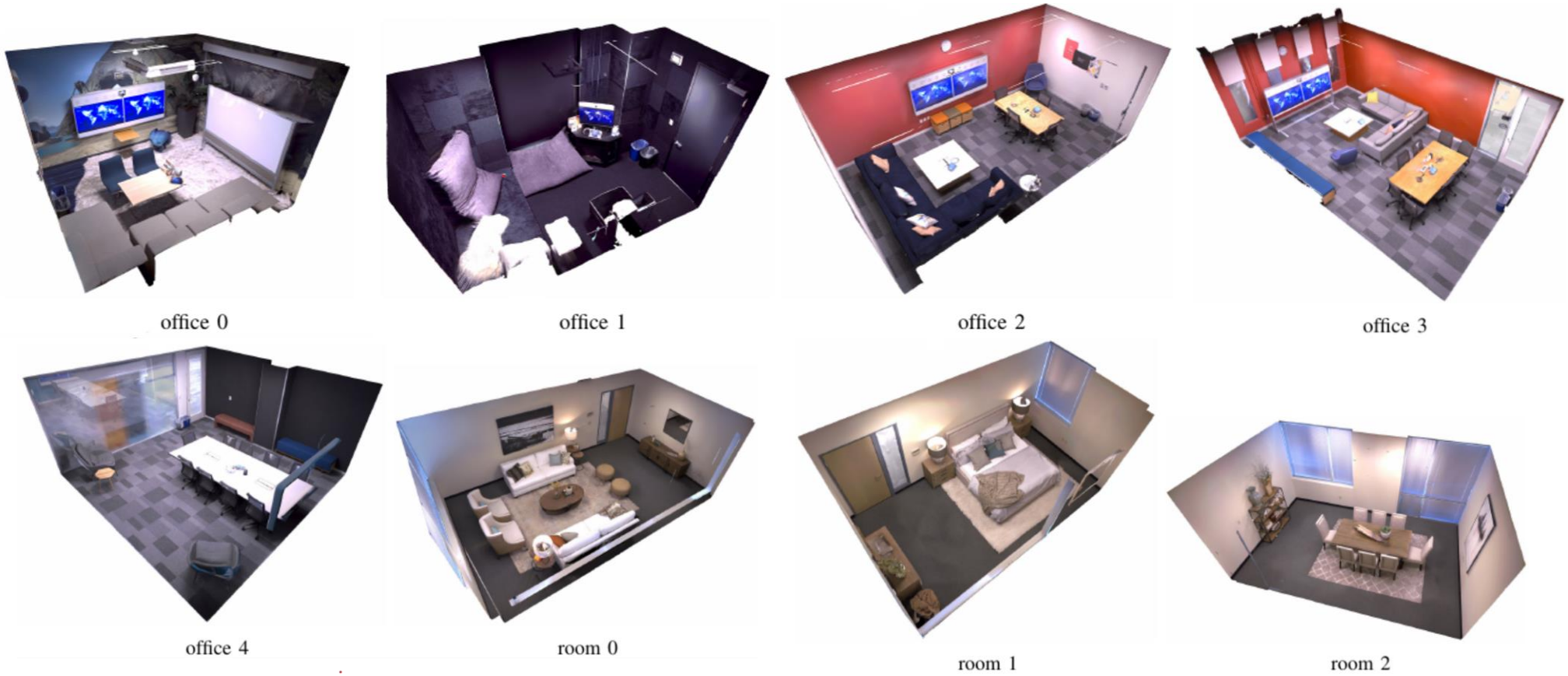
Optimization Using Event Loss

- Finally, the event loss is backpropagated through differentiable Event-Net and renderer to optimize camera poses
- Pretrained parameters of Event-Net fixed during the entire run



Experiments: Setup

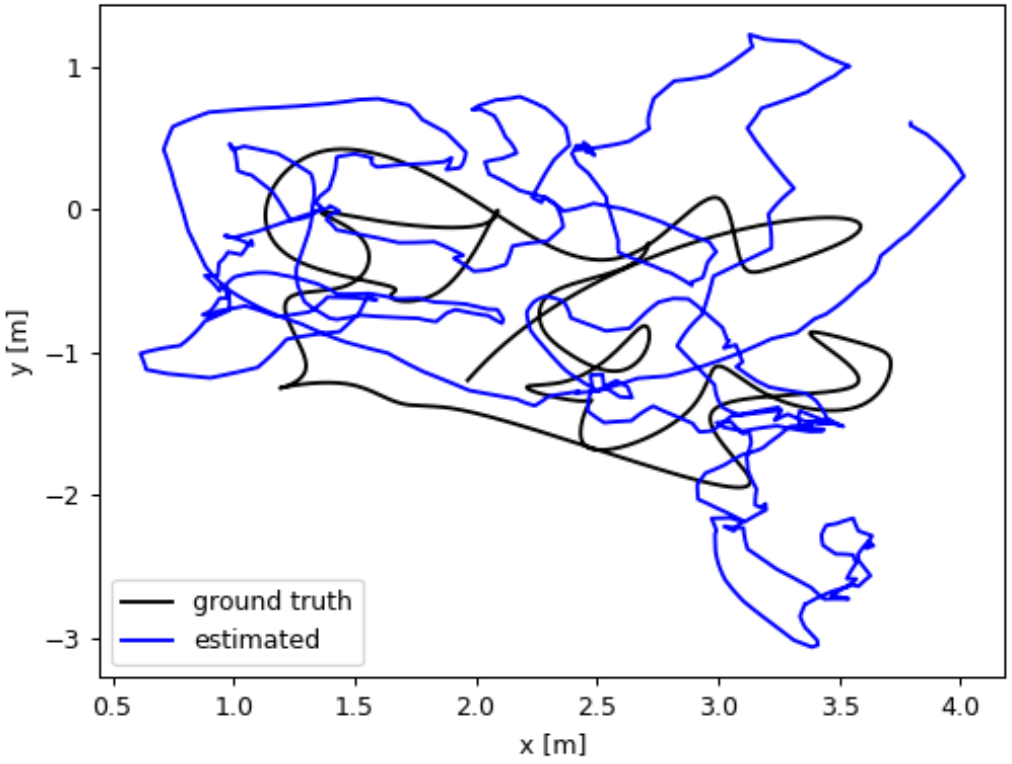
- **Dataset**: Replica RGB-D image sequences rendered in 8 scenes, each 2000 frames long
- **Comparison**:
  - NICE-SLAM: fed with RGB-D every fifth frame (simulates faster camera motion)
  - **EvenNICER-SLAM**: RGB-D available every fifth frame, event images always available



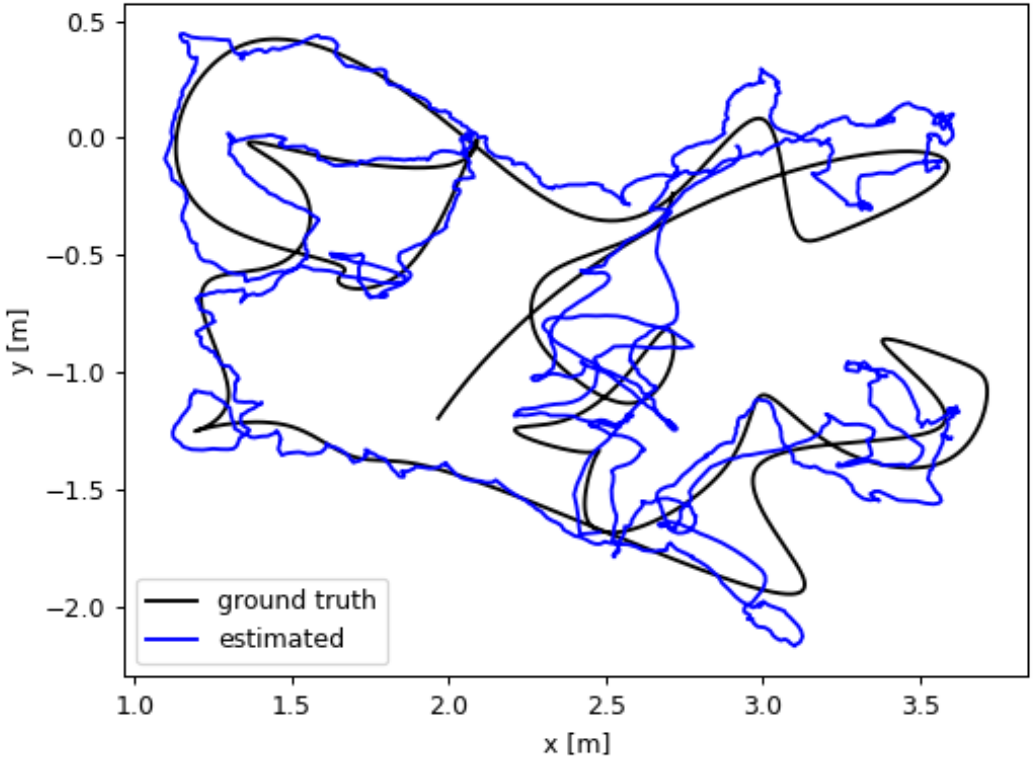
(https://github.com/facebookresearch/Replica-Dataset)

Qualitative Results: Camera Tracking

- EvenNICER-SLAM keeps better track of the camera



NICE-SLAM



EvenNICER-SLAM

Quantitative Results: Camera Tracking

- NICE-SLAM loses track of the camera easily with sparse input!
- EvenNICER-SLAM significantly reduces failure cases and tracking error
- Metric: ATE RMSE [cm], the smaller the better

Scene	room0	room1	room2	office0	office1	office2	office3	office4
NICE-SLAM [25]	F(701)	F(380)	262.6	135.0	F(350)	F(1754)	F(658)	F(465)
Ours	43.6	89.1	19.2	42.1	67.9	50.5	42.6	F(1352)

Qualitative Results: Mapping

(a) Ground truth

(b) NICE-SLAM [25]

(c) **EvenNICER-SLAM** (ours)

Quantitative Results: Mapping

- Similar tendency

	Metric	room0	room1	room2	office0	office1	office2	office3	office4
NICE-SLAM [25]	Depth L1	F	F	86.79	78.74	F	F	F	F
	Accuracy	F	F	149.92	107.19	F	F	F	F
	Completion	F	F	22.90	29.78	F	F	F	F
	Comp. Ratio	F	F	18.35	17.56	F	F	F	F
Ours	Depth L1	73.55	84.22	11.95	40.93	44.09	22.94	36.79	F
	Accuracy	50.96	71.70	8.31	43.73	37.96	13.86	25.31	F
	Completion	39.79	40.20	7.50	24.97	26.21	11.24	20.79	F
	Comp. Ratio	21.46	13.62	52.33	16.33	14.68	33.53	26.93	F

Discussion: What do the results suggest?

- The event supervision can positively contribute to camera tracking
- Also indirectly improves mapping quality

Still Room for Improvement!

- Still not as good as NICE-SLAM with full input
  - Not a fair comparison because an event image essentially provides less information than a set of RGB-D images do
  - But we believe EvenNICER-SLAM can be improved to get similar performance

(c) **EvenNICER-SLAM** (ours)

(d) NICE-SLAM [25] (full input)

Future Work

- More tractable alternative for event loss: grayscale photometric loss
  - Next grayscale image ≈ current grayscale image + event image * contrast threshold
  - Direct supervision of rendered images using photometric loss
  - Expected to solve alignment issue, more stable optimization

Future Work

- More tractable alternative for event loss: grayscale photometric loss
  - Next grayscale image ≈ current grayscale image + event image * contrast threshold
  - Direct supervision of rendered images using photometric loss
  - Expected to solve alignment issue, more stable optimization
- Event Loss for Mapping: event → optical flow → mapping
  - Event → optical flow: E-RAFT [Gehrig2021]
  - Optical flow → mapping: DROID-SLAM [Teed2021], NICER-SLAM [Zhu,Peng2023]
  - Possible to get rid of depth map input → Reduces data collection cost

Future Work

- More tractable alternative for event loss: grayscale photometric loss
  - Next grayscale image ≈ current grayscale image + event image * contrast threshold
  - Direct supervision of rendered images using photometric loss
  - Expected to solve alignment issue, more stable optimization
- Event Loss for Mapping: event → optical flow → mapping
  - Event → optical flow: E-RAFT [Gehrig2021]
  - Optical flow → mapping: DROID-SLAM [Teed2021], NICER-SLAM [Zhu,Peng2023]
  - Possible to get rid of depth map input → Reduces data collection cost
- EvenNICER-SLAM in the real domain?
  - Low-latency & no-motion-blur event cameras are good
  - Real-synthetic domain gap: Real event data can be noisy → Adversarial training of Event-Net
  - Potential for high-speed applications: autonomous drones, self-driving vehicles…

References

- Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., ... & Pollefeys, M. (2022). Nice-slam: Neural implicit scalable encoding for slam. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12786-12796).
- Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., ... & Newcombe, R. (2019). The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797.
- Hanover, D., Loquercio, A., Bauersfeld, L., Romero, A., Penicka, R., Song, Y., ... & Scaramuzza, D. (2023). Autonomous Drone Racing: A Survey. arXiv e-prints, arXiv-2301.
- Guan, Y., Hou, X., Wu, N., Han, B., & Han, T. (2022, June). DeepMix: mobility-aware, lightweight, and hybrid 3D object detection for headsets. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (pp. 28-41).
- Rebecq, H., Ranftl, R., Koltun, V., & Scaramuzza, D. (2019). Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3857-3866).
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.
- Gehrig, M., Millhäusler, M., Gehrig, D., & Scaramuzza, D. (2021, December). E-raft: Dense optical flow from event cameras. In 2021 International Conference on 3D Vision (3DV) (pp. 197-206). IEEE.
- Teed, Z., & Deng, J. (2021). Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems, 34, 16558-16569.
- Zhu, Z., Peng, S., Larsson, V., Cui, Z., Oswald, M. R., Geiger, A., & Pollefeys, M. (2023). NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM. arXiv preprint arXiv:2302.03594.

ETH zürich

Thanks for listening!