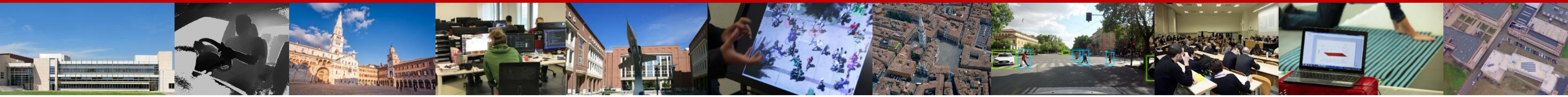
VISAPP 2020
15th International Conference on Computer Vision Theory and Applications
Valletta - Malta 27 - 29 February, 2020
VISIGRAPP

Learn to See by Events: Color Frame Synthesis from Event and RGB Cameras



Stefano Pini, Guido Borghi, Roberto Vezzani

s.pini@unimore.it, guido.borghi@unimore.it, roberto.vezzani@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

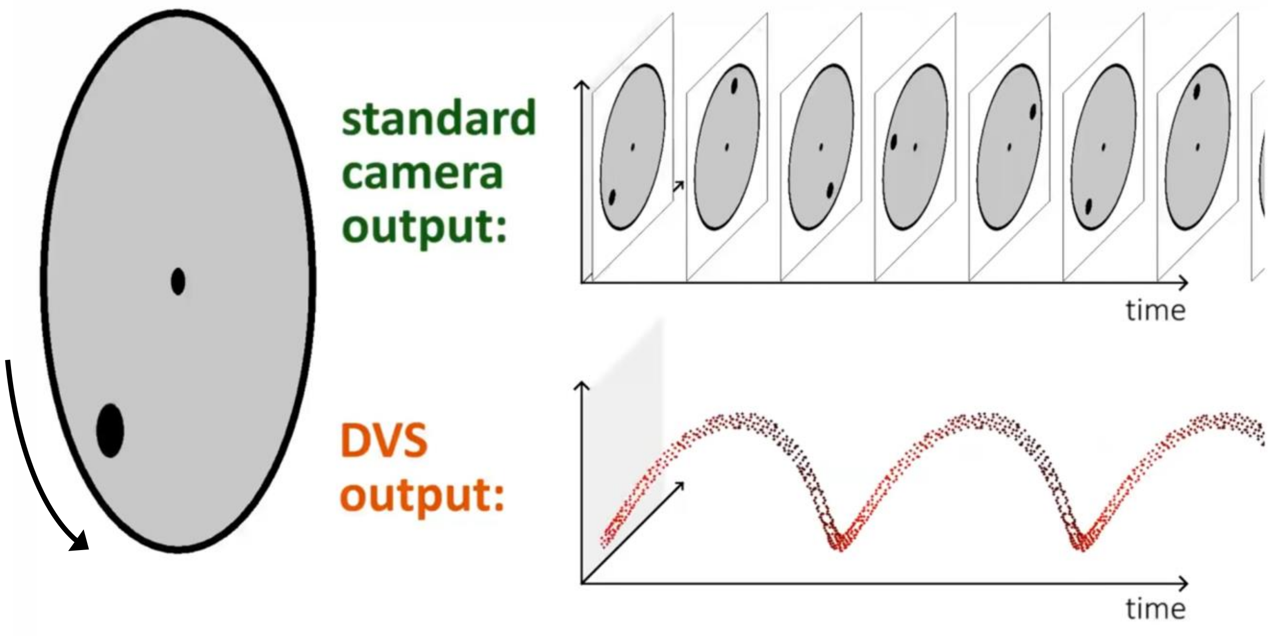
Centro Interdipartimentale di Ricerca Softech: ICT per le Imprese

Outline

- Introduction to event cameras
- Contributions
- Mathematical Formulation
- Proposed method
- Experimental evaluation - Datasets and Metrics
- Experimental evaluation - Results
- Conclusions

Event Cameras

- Event cameras are **biologically-inspired sensors** that gather the temporal evolution of the scene.
- They **capture pixel-wise brightness variations** and output a stream of **asynchronous** events
- The major advantages of this type of neuromorphic sensors are:
  - **Low power consumption**
  - **Low data rate**
  - **High temporal resolution**
  - **High dynamic range**

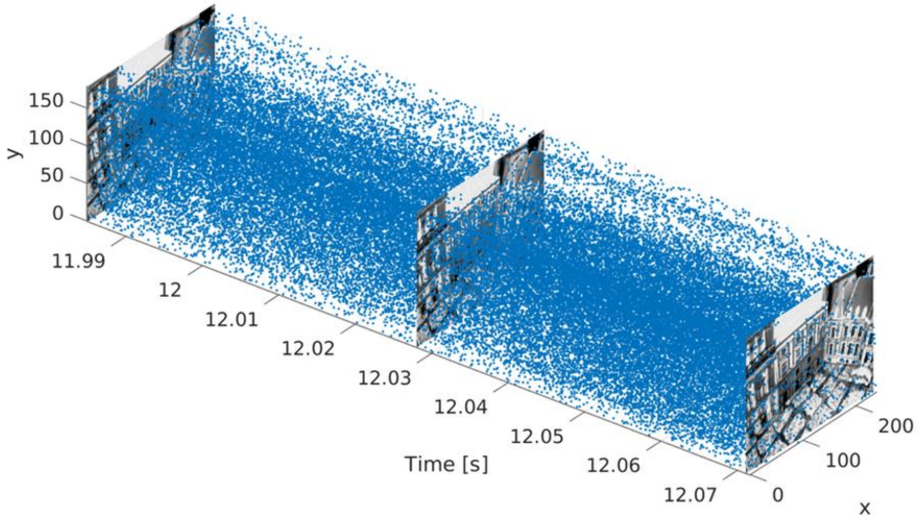


1. *Video credits: https://youtu.be/LauQ6LWTkxM (Mueggler et al., IROS 2014)*

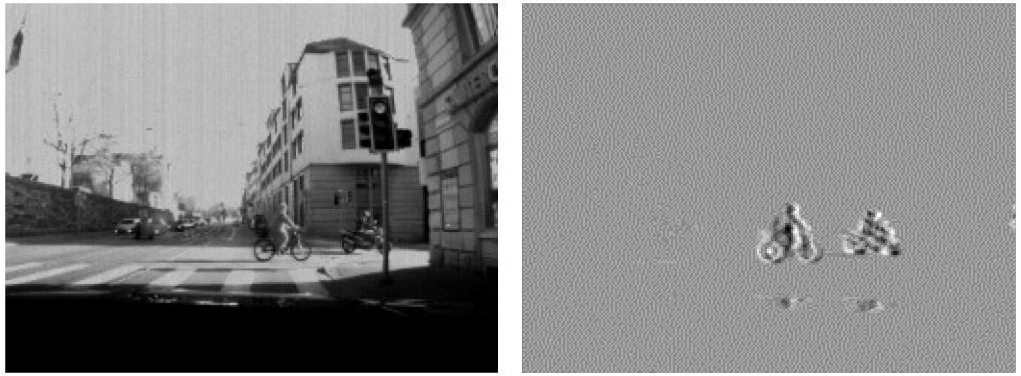
Limitations

- Despite having multiple advantages with respect to traditional cameras, their practical use is partially limited because:
  - limited applicability of traditional data processing
  - limited applicability of traditional vision algorithms
  - need to acquire new datasets

- **Do we really need to reinvent the wheel?**



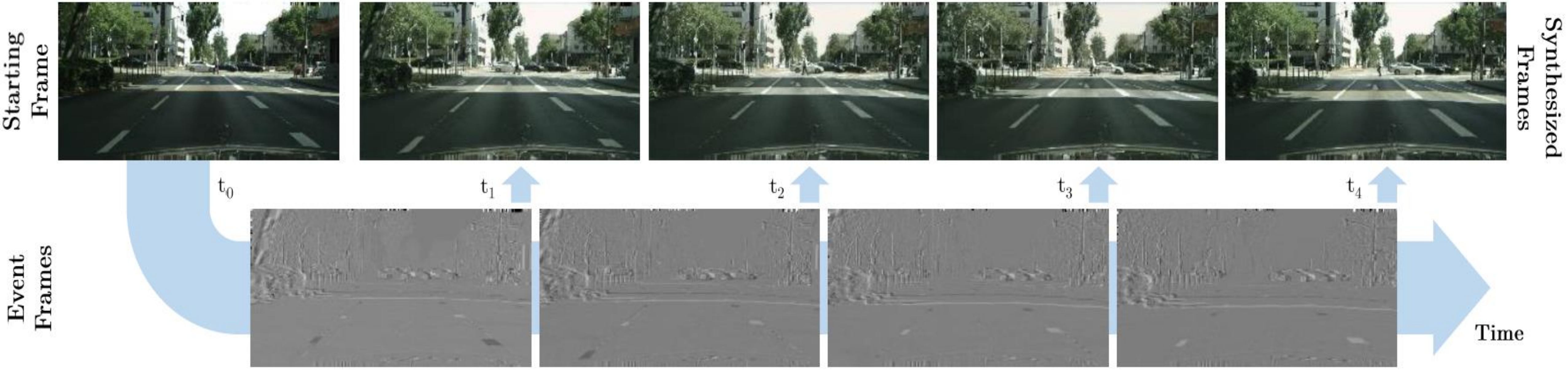
Events through time



Gray-level frame and triggered events.
Where brightness variations don't occur, no events are triggered!

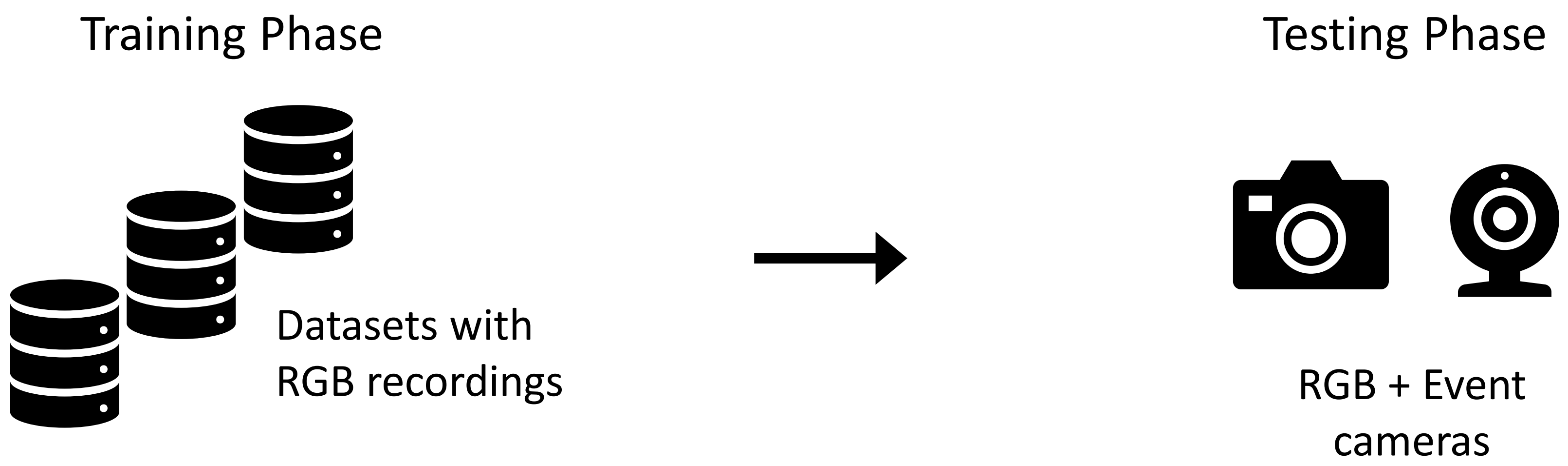
Contributions (1)

- We propose a deep learning-based method that **synthesizes frames** using asynchronous events from an event camera and RGB key-frames from a low-framerate camera
  - The method **increases the temporal resolution** of the RGB stream **preserving high-quality textures and details**
  - Traditional **vision algorithms can be directly applied** on synthesized frames



Contributions (2)

- We investigate the of use **simulated event data** from RGB videos so that
  - Event-based methods can be **evaluated on standard annotated datasets**, which are often not available in the event domain.
  - Learned models can be **trained on simulated event data** and used with real event data, unseen during the training procedure



Mathematical Formulation (1)

Event

a positive/negative brightness variation at position (x,y) and time t

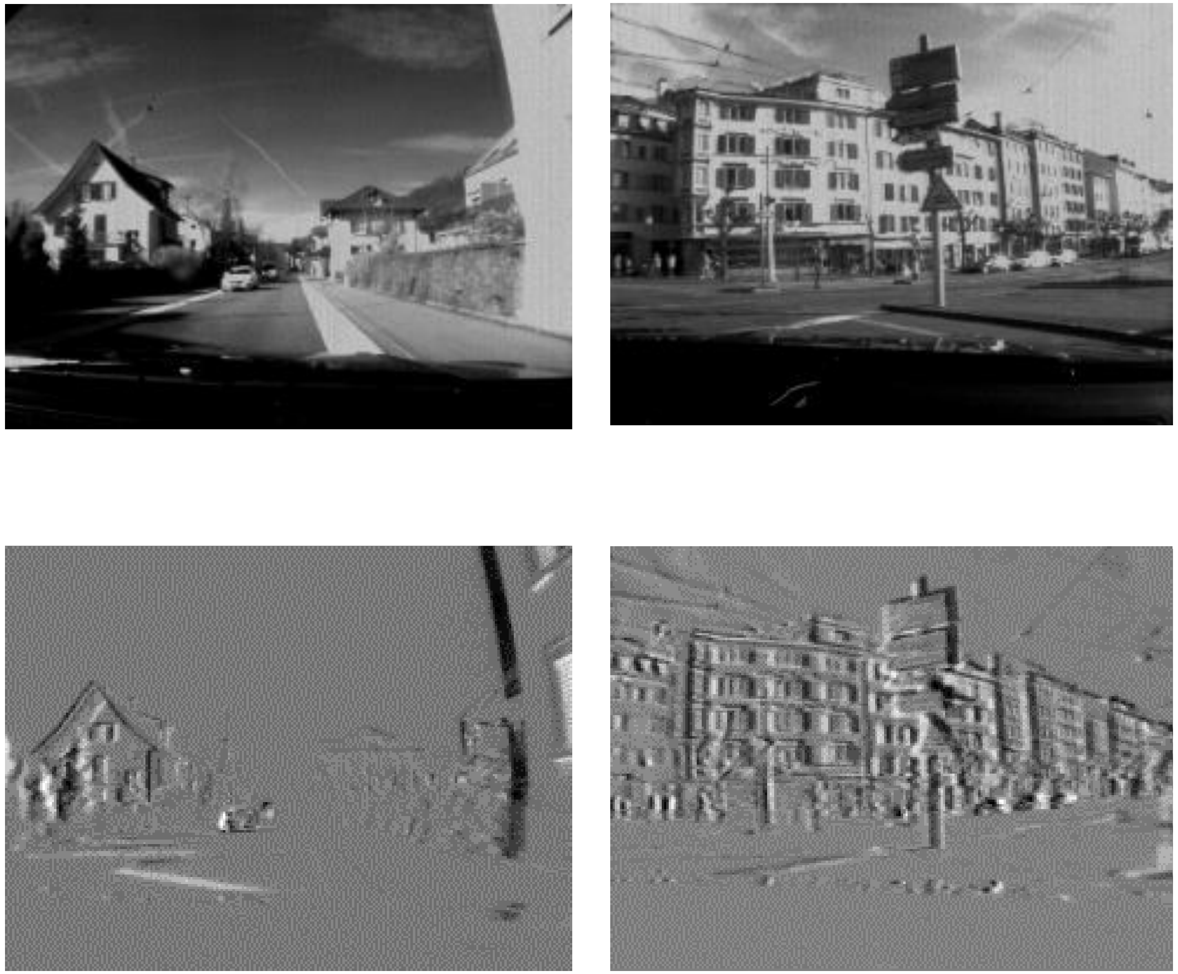
$$e_k = (x_k, y_k, t_k, p_k)$$

Event Frame

a pixel-wise sum of all events occurred in a time interval

$$\Phi_\tau(t) = \sum_{e_k \in E_{t,\tau}} p_k$$

$$E_{t,\tau} = \{e_k \,|\, t_k \in [t,\ t+\tau]\}$$



1. *Maqueda et al., "Event-based vision meets deep learning on steering prediction for self-driving cars". CVPR, 2018.*

Mathematical Formulation (2)

Brightness variation approximation

for small time intervals, the brightness variations can be approximated with a First-order Taylor approximation

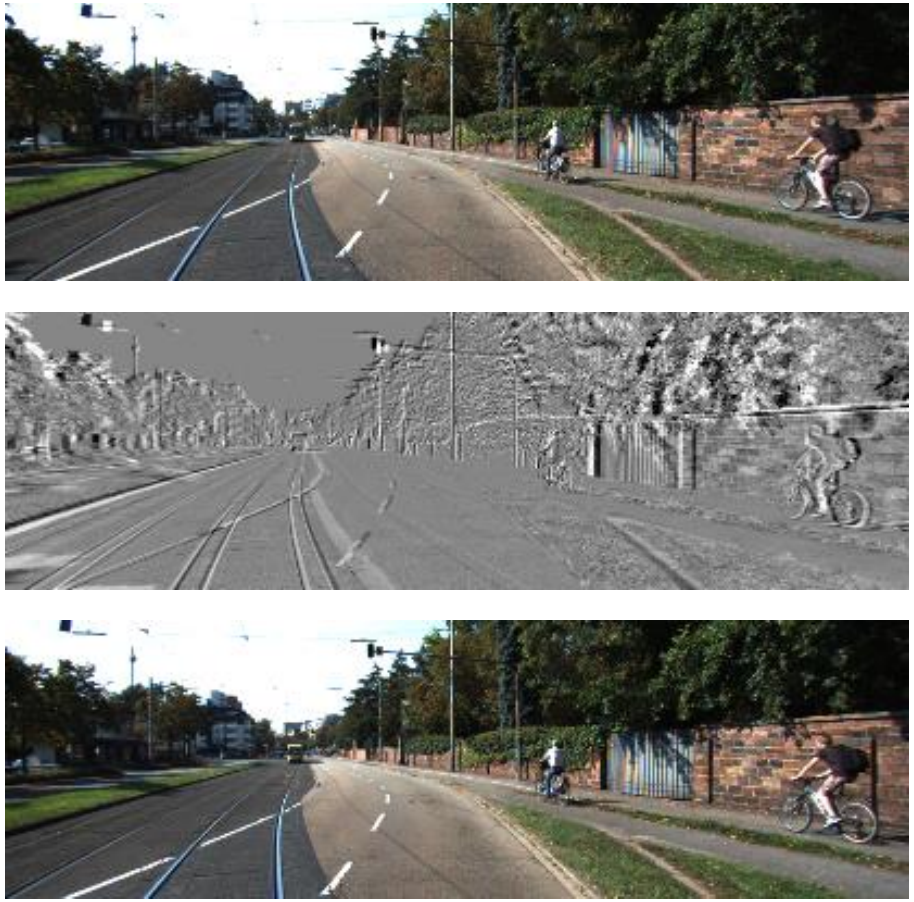
$$\lim_{\tau \to 0} \frac{\delta L}{\delta t} \tau \approx L(t+\tau) - L(t) \doteq \Delta L$$

$$L(t) = \log\left(Br\left(I(t)\right)\right)$$

Event frame approximation

a synthetic approximation of an event frame can be obtained subtracting the log-brightness of two standard frames

$$\Phi_\tau(t) \approx \Delta L = \log\left[Br\left(I(t+\tau)\right)\right] - \log\left[Br\left(I(t)\right)\right]$$



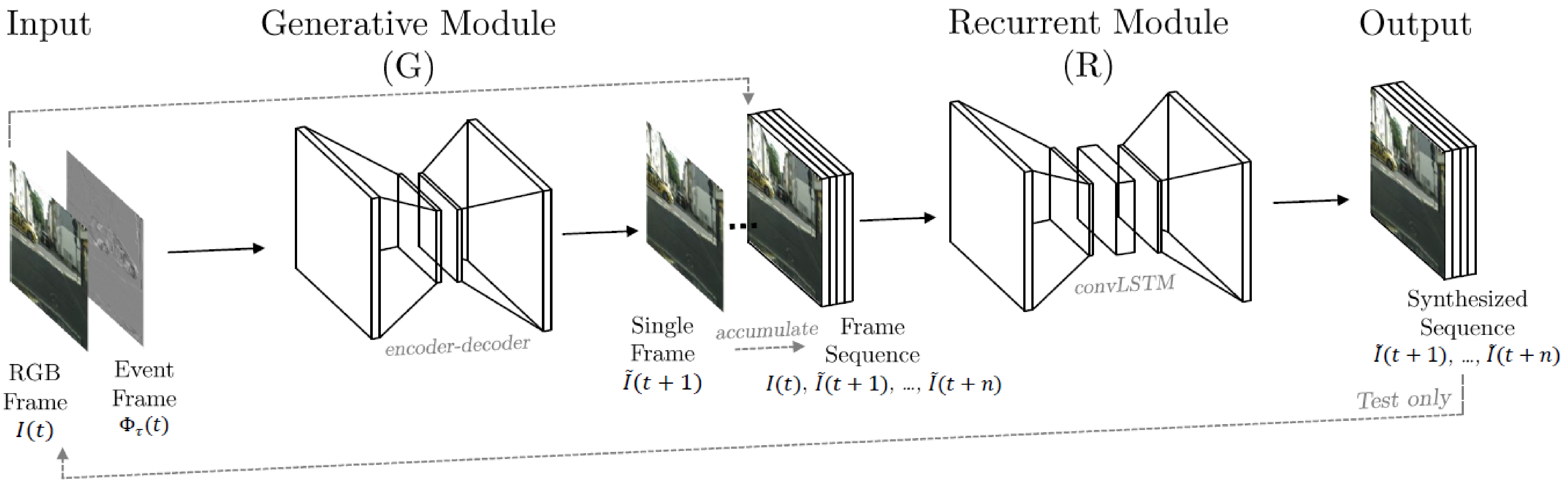
1. *Maqueda et al., “Event-based vision meets deep learning on steering prediction for self-driving cars”. CVPR, 2018.*
2. *Gehrig et al., “Asynchronous, photometric feature tracking using events and frames”. ECCV, 2018.*

Proposed Method (1)

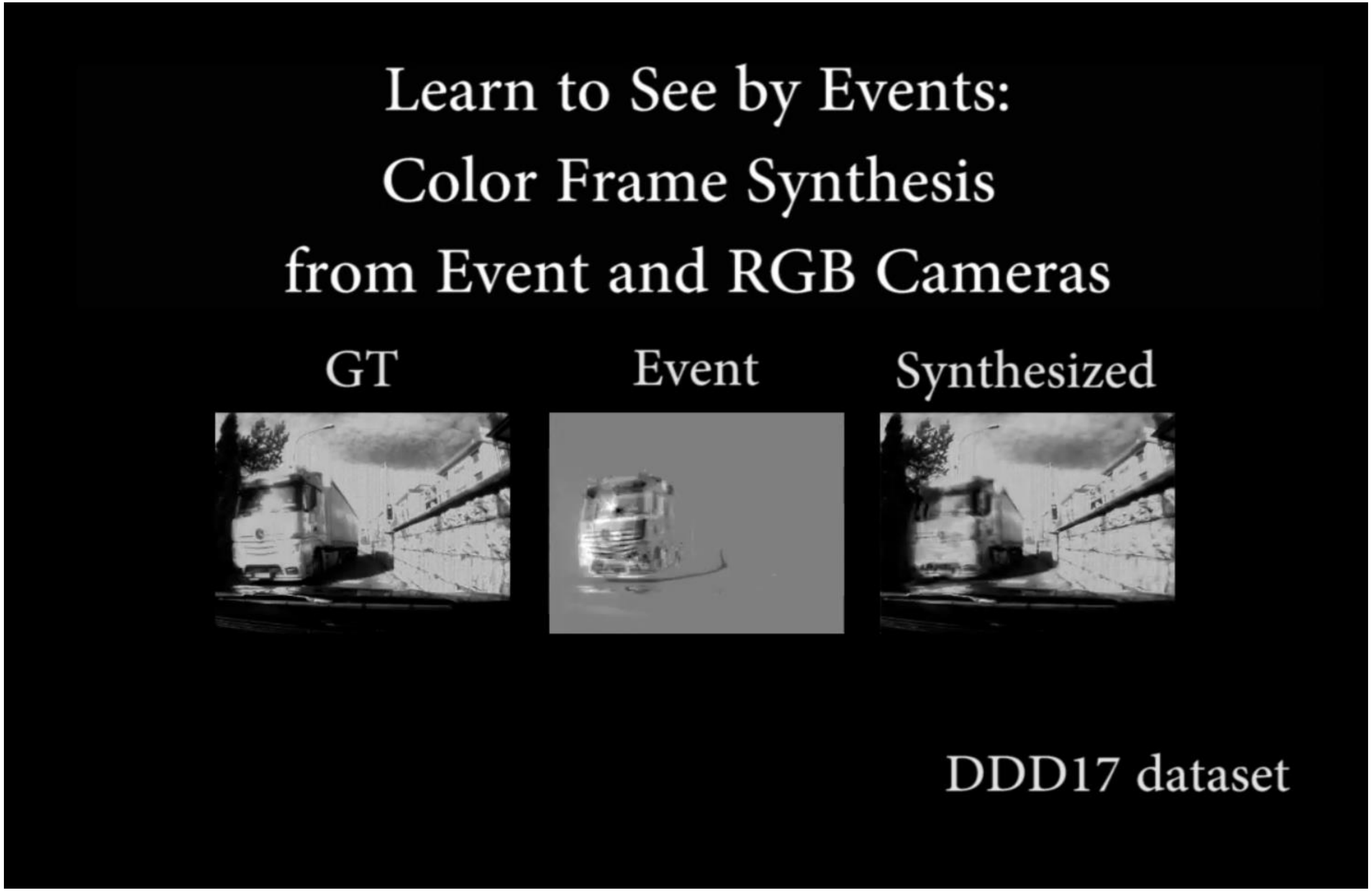
- We propose a deep generative architecture composed of two main modules:
  - **Generative Module (G)** – next frame generation
    - Input: RGB frame + event frame
    - Output: next RGB frame
    - Architecture: U-Net
    - Loss: L2 + adversarial
  - **Recurrent Module (R)** – temporal sequence refinement
    - Input: RGB key-frame + N generated RGB frame
    - Output: N generated RGB frames
    - Architecture: U-Net-like architecture with a convLSTM in the bottleneck
    - Loss: L2

Proposed Method (2)

Overall architecture



Visual Results



1. https://youtu.be/vC2dGc88tq4

Experimental Evaluation - Datasets

- We employed four automotive datasets

Two event datasets

DDD17

MVSEC

Two famous RGB datasets

Kitti

Cityscapes

Experimental Evaluation – Pixel-wise and Perceptual Metrics

Pixel-wise metrics

L1, L2, RMSE

Thresholds

$$J_i = \{y \in I \,|\, max(\frac{y}{\hat{y}}, \frac{\hat{y}}{y}) < 1.25^i\} \qquad \delta_i = \frac{1}{|I|}\,|J_i|$$

PSNR

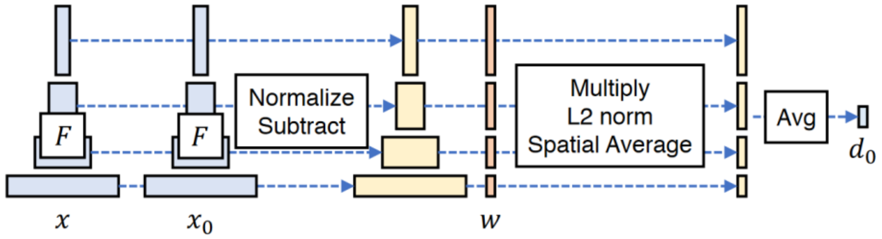
$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{m}{L_2}\right)$$

SSIM[1]

$$\mathrm{SSIM}(p,q) = \frac{(2\mu_p\mu_q + c_1)(2\sigma_{pq} + c_2)}{(\mu_p^2 + \mu_q^2 + c_1)(\sigma_p^2 + \sigma_q^2 + c_2)}$$

Perceptual metrics

LPIPS[2]



1. Wang et al. "Image quality assessment: from error visibility to structural similarity". IEEE transactions on image processing, 2004.
2. Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". CVPR, 2018.

Experimental Evaluation – Results (1)

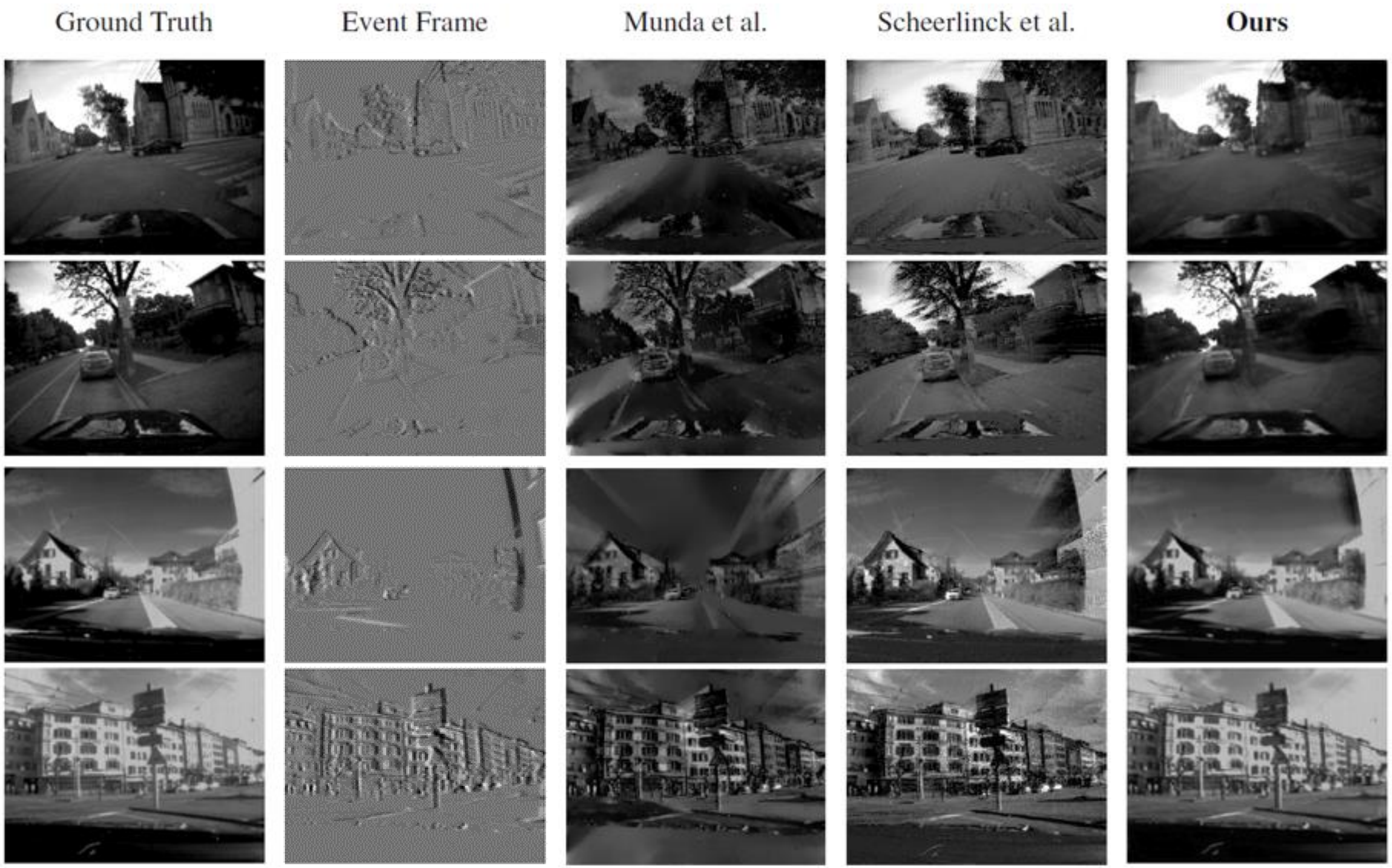
Results on **real event datasets** (DDD17[1], MVSEC[2])

Comparison between our framework and the literature

Dataset	Method	Norm ↓ L_1	Norm ↓ L_2	RMSE ↓ Lin	RMSE ↓ Log	RMSE ↓ Scl	Threshold ↑ 1.25	Threshold ↑ 1.25^2	Threshold ↑ 1.25^3	Indexes ↑ PSNR	Indexes ↑ SSIM	Perceptual ↓ LPIPS
DDD17	Munda et al.	0.268	94.277	0.314	5.674	5.142	0.152	0.448	0.536	10.244	0.216	0.637
	Scheerlinck et al.	0.080	29.249	0.098	4.830	4.352	0.671	0.781	0.827	20.542	0.702	0.208
	Pini et al.	0.027	8.916	0.040	4.048	3.571	0.775	0.848	0.875	29.176	0.864	**0.105**
	Ours	**0.022**	**8.583**	**0.039**	**3.766**	**3.408**	**0.787**	**0.855**	**0.880**	**29.428**	**0.884**	0.107
MVSEC	Munda et al.	0.160	86.419	0.288	8.985	8.016	0.088	0.163	0.232	11.034	0.181	0.599
	Scheerlinck et al.	0.067	26.794	0.089	7.313	6.982	0.263	0.357	0.467	21.070	0.551	0.257
	Pini et al.	0.026	12.062	0.054	**6.443**	6.102	**0.525**	**0.642**	**0.708**	25.866	0.740	0.172
	Ours	**0.022**	**11.216**	**0.051**	6.559	**6.003**	0.514	0.637	0.699	**26.366**	**0.845**	**0.137**

1. *Binas et al. "Ddd17: End-to-end davis driving dataset". ICML Workshop on Machine Learning for Autonomous Vehicles (MLAV), 2017.*
2. *Zhu et al. "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception". IEEE Robotics and Automation Letters, 2018.*
3. *Munda et al. "Real-time intensity-image reconstruction for event cameras using manifold regularisation". IJCV, 2018.*
4. *Scheerlinck et al. "Continuous-time intensity estimation using event cameras". ACCV, 2018.*
5. *Pini et al. "Video synthesis from intensity and event frames". ICIAP, 2019.*

Experimental Evaluation - Qualitative Results (1)



Experimental Evaluation – Results (2)

Results on **real event datasets** (DDD17[1], MVSEC[2]) and **synthetic event datasets** (KITTI[3], Cityscapes[4])

Comparison between our whole framework (**G**enerative+**R**ecurrent) and the **G**enerative module

Dataset	Model	Norm ↓		Difference ↓		RMSE ↓			Threshold ↑			Indexes↑	
		L_1	L_2	Abs	Sqr	Lin	Log	Scl	1.25	1.25^2	1.25^3	PSNR	SSIM
DDD17	G	0.029	9.658	0.114	0.007	0.044	2.296	2.268	0.854	0.919	0.941	28.486	0.876
	G+R	0.022	8.583	0.167	0.006	0.039	3.766	3.408	0.787	0.855	0.880	29.428	0.884
MVSEC	G	0.026	12.830	0.311	0.013	0.058	6.302	6.233	0.562	0.675	0.733	25.309	0.784
	G+R	0.022	11.216	0.354	0.010	0.051	6.559	6.003	0.514	0.637	0.699	26.366	0.845
Kitti	G	0.030	10.95	0.125	0.006	0.048	0.472	0.463	0.782	0.940	0.981	27.140	0.919
	G+R	0.029	10.71	0.105	0.005	0.046	0.194	0.191	0.846	0.968	0.991	27.295	0.928
CS	G	0.019	4.534	0.086	0.003	0.025	0.232	0.211	0.877	0.974	0.992	32.769	0.962
	G+R	0.015	4.192	0.059	0.002	0.023	0.172	0.170	0.968	0.997	0.999	33.315	0.971

1. *Binas et al. "Ddd17: End-to-end davis driving dataset". ICML Workshop on Machine Learning for Autonomous Vehicles (MLAV), 2017.*
2. *Zhu et al. "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception". IEEE Robotics and Automation Letters, 2018.*
3. *Geiger et al. "Vision meets robotics: The kitti dataset". International Journal of Robotics Research (IJRR), 2013.*
4. *Cordts et al. "The cityscapes dataset for semantic urban scene understanding". CVPR, 2016.*

AImage Lab

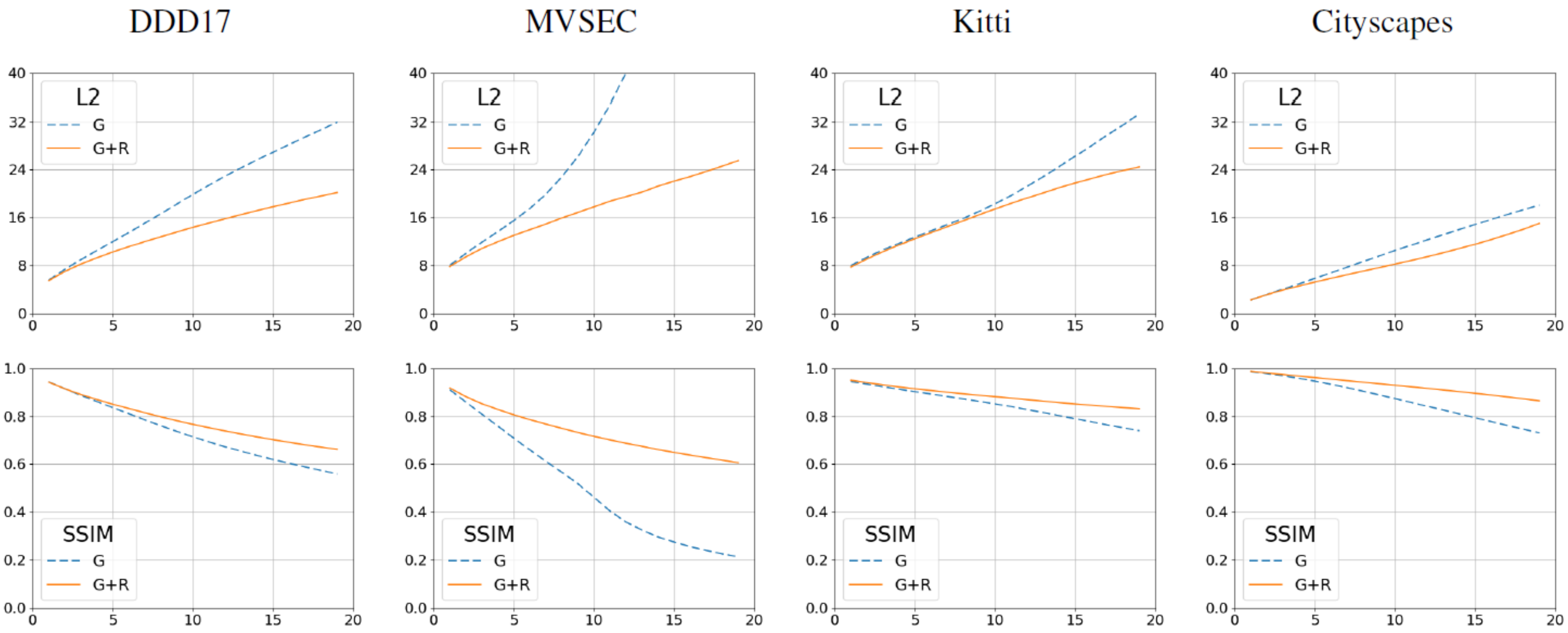
Experimental Evaluation - Qualitative Results (2)

G.T.	Generative Module	Recurrent Module

Experimental Evaluation – Results (3)

Variation of **L2** and **SSIM** as a function of the number of subsequently-synthesized frames after the last key-frame.

The horizontal axis refers to the frame on which the metric is calculated, starting from an initial color frame and estimating the following ones.



Experimental Evaluation – Task-based Metrics

Semantic Segmentation

- We adopt a **pre-trained semantic classifier** (WideResNet+38+DeepLab3[1], trained on Cityscapes) to measure the accuracy of a certain set of pixels to be a particular class.

Object detection

- We adopt a **pre-trained object detector** (Yolo-v3[2], trained on COCO[3]) to investigate the ability of the proposed framework to preserve objects in the generated frames, in particular people, trucks, cars, buses, trains, and stop signals.

Underlying Idea

If synthesized images are close to the real ones → the classifier/detector will achieve good results

1. Rota Bulò et al. *"In-place activated batchnorm for memory-optimized training of dnns". CVPR, 2018.*
2. Redmonet al. *"YOLOv3: An incremental improvement.", arXiv preprint, 2018.*
3. Lin et al. *"Microsoft COCO: Common objects in context". ECCV, 2014.*

Experimental Evaluation - Results (4)

Semantic Segmentation and **Object Detection** scores on synthesized frames from **Kitti**[1] and **Cityscapes**[2].

Comparison between the Generative module (**G**), the whole framework (**G+R**), and the Ground Truth (**GT**).

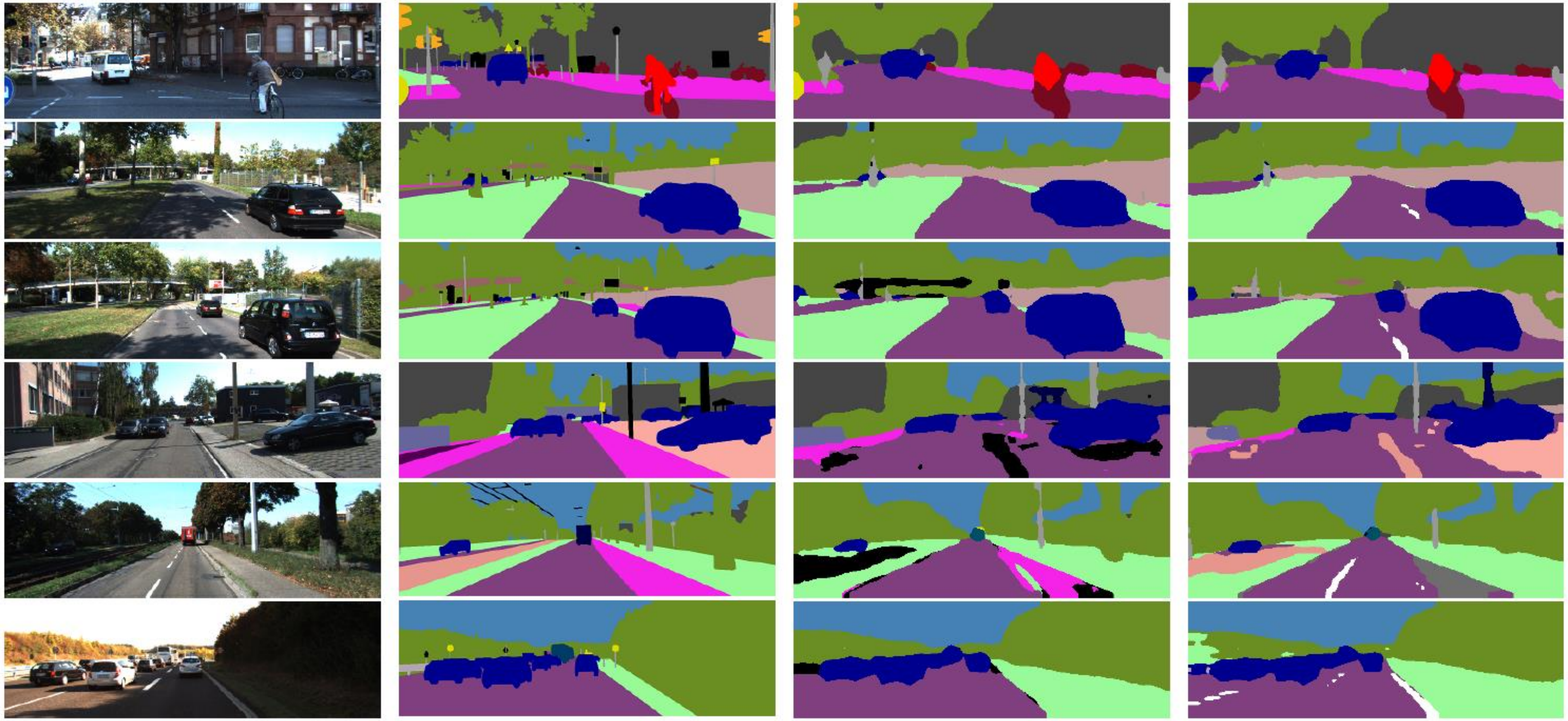
Data	**Model**	**Semantic Segmentation** ↑ Per-pixel	Per-class	class IoU	**Object Det.** ↑ mIoU	%
Kitti	G	0.814	0.261	0.215	0.914	65.8
	G+R	0.813	0.261	0.215	0.912	71.4
	GT	0.827	0.283	0.235	-	-
CS	G	0.771	0.197	0.162	0.924	83.5
	G+R	0.790	0.201	0.166	0.926	86.4
	GT	0.828	0.227	0.192	-	-

1. *Geiger et al. "Vision meets robotics: The kitti dataset". International Journal of Robotics Research (IJRR), 2013.*
2. *Cordts et al. "The cityscapes dataset for semantic urban scene understanding". CVPR, 2016.*

Experimental Evaluation - Qualitative Results (4)

Semantic Segmentation on Kitti dataset

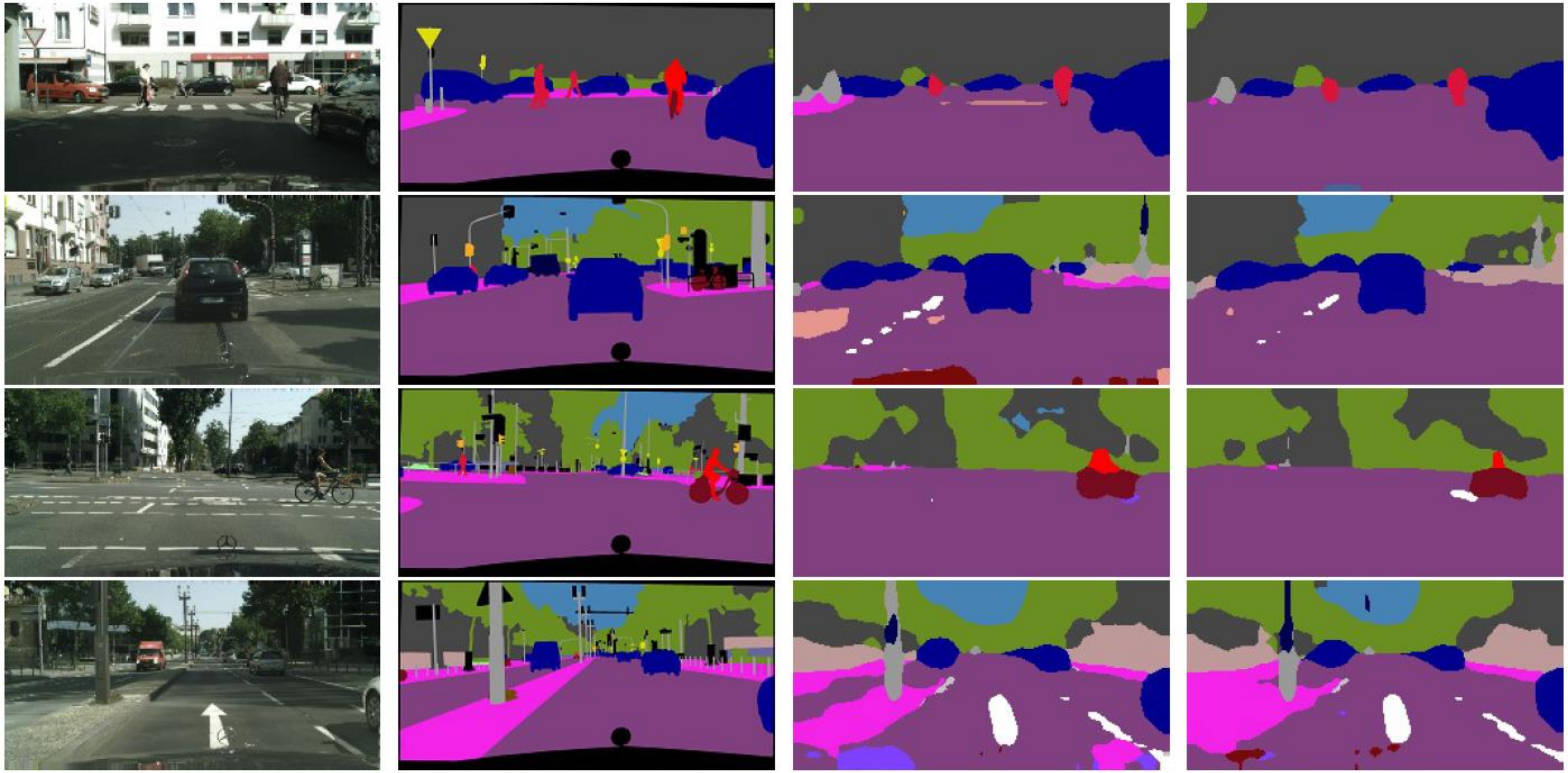
Original Frame | G.T. | Original Frame | Synth. Frame



Experimental Evaluation - Qualitative Results (4)

Semantic Segmentation on Cityscapes dataset

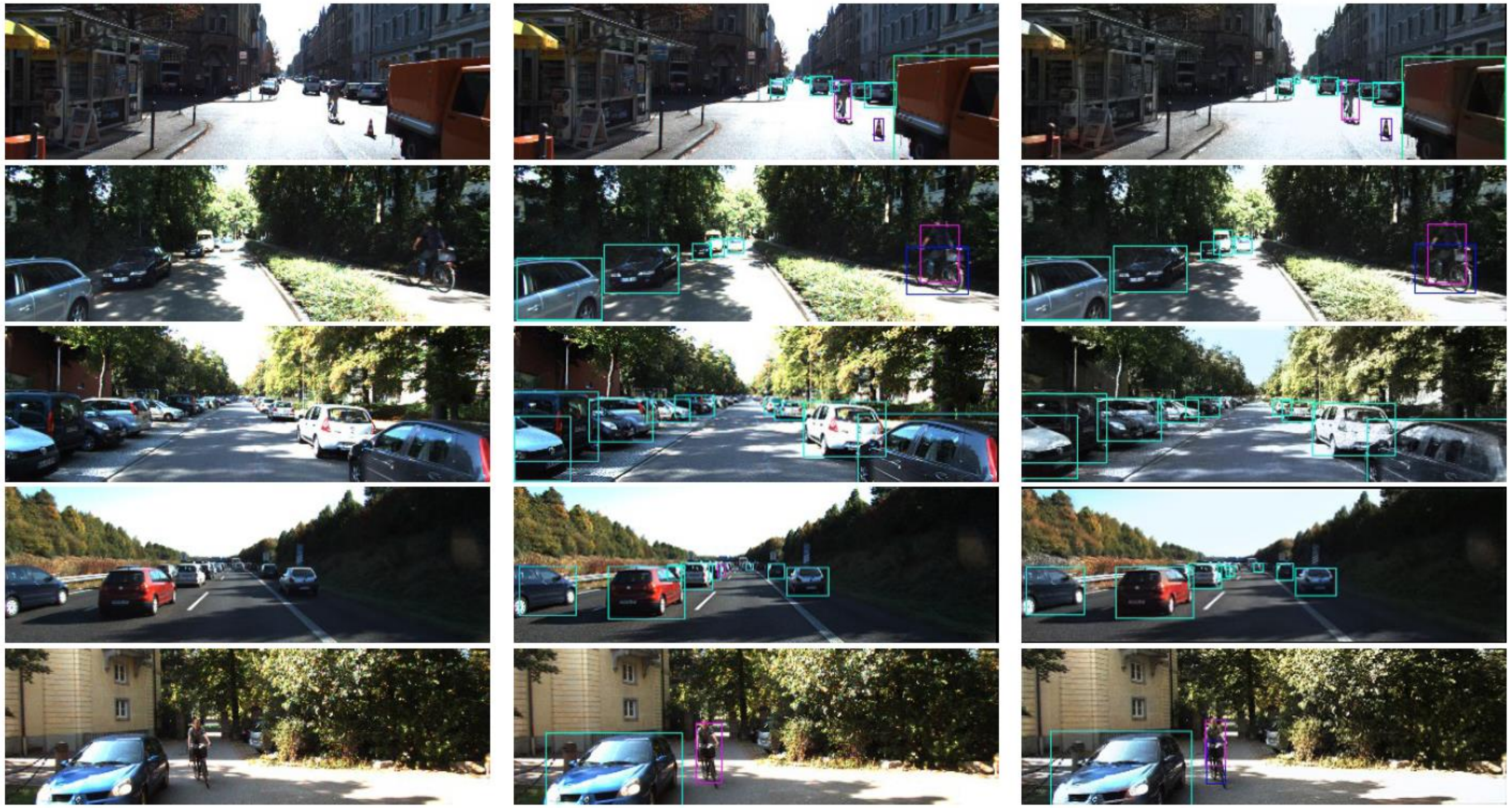
Original Frame	G.T.	Original Frame	Synth. Frame



Experimental Evaluation - Qualitative Results (4)

Object detection on Kitti

Original Frame | G.T. | Synthesized Frame



Experimental Evaluation - Qualitative Results (4)

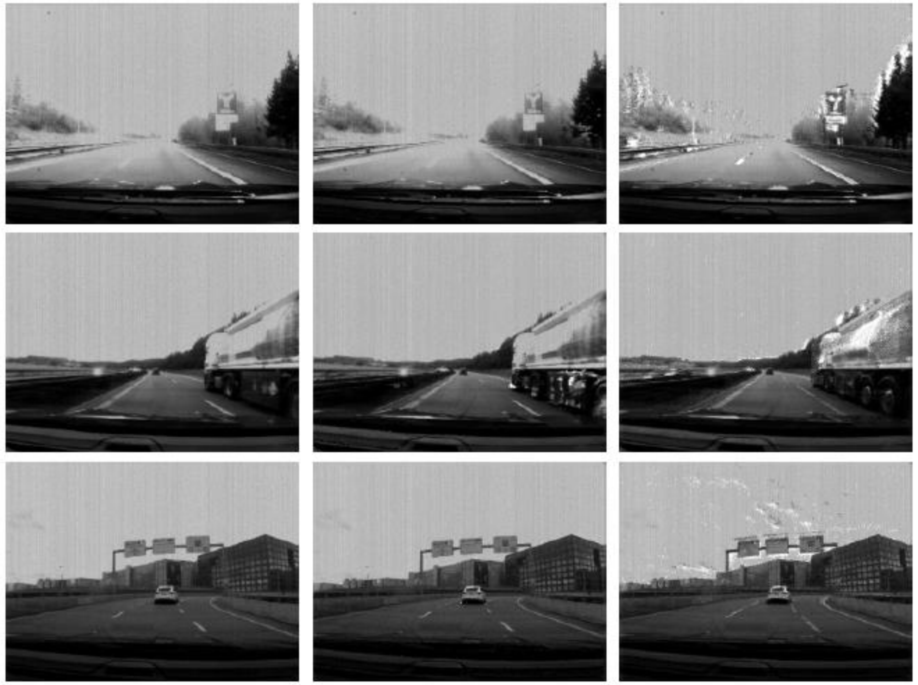
Object detection on Cityscapes

Original Frame G.T. Synthesized Frame

Experimental Evaluation - Cross-modality Test

- We **train** our model on DDD17 and MVSEC datasets, using **simulated event frames**.
- We **test** the network using as input **real event frames**, without any fine-tuning procedure.
- On DDD17, we obtain PSNR of 23.396 and SSIM of 0.779. (~15% drop)
- On MVSEC, we obtain PSNR of 21.935 and SSIM of 0.736. (~15% drop)

G.T. | Synth. events | Real events

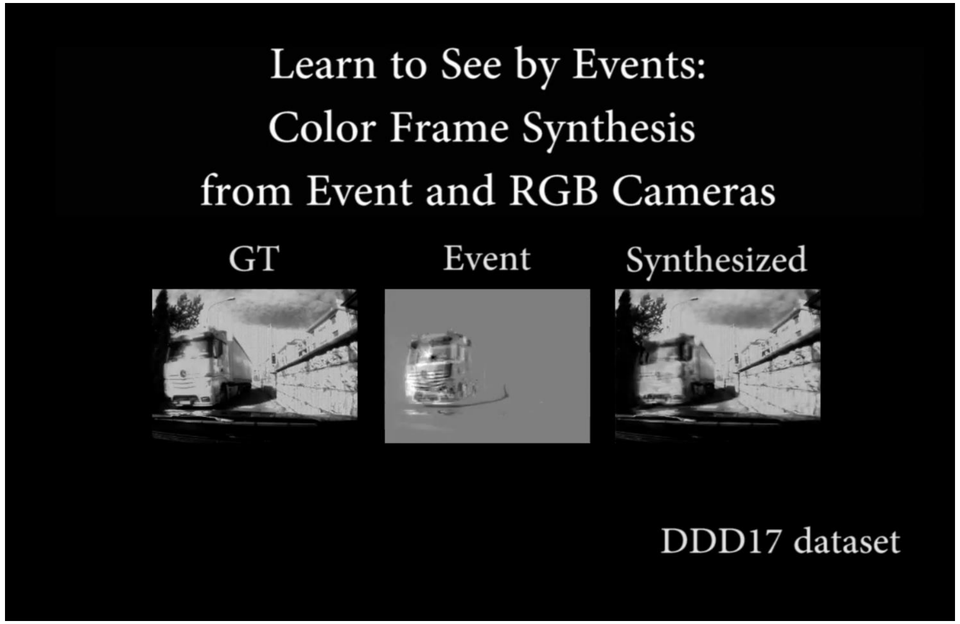


G.T. | Synth. events | Real events

Conclusions

- We proposed a framework that **synthesizes color frames**, relying on an initial or a periodic set of key-frames and a **sequence of event frames**.
  - The method **preserve high-quality textures and details**
  - Traditional **vision algorithms can be successfully applied** on synthesized frames

- We used **simulated event frames** and **evaluate the accuracy of our method on standard annotated datasets**, which are not available in the event domain.



1. Video: *https://youtu.be/vC2dGc88tq4*

VISAPP 2020

15th International Conference on Computer Vision Theory and Applications

Valletta - Malta 27 - 29 February, 2020

VISIGRAPP

Thank you for your attention

Learn to See by Events: Color Frame Synthesis from Event and RGB Cameras

Stefano Pini, Guido Borghi, Roberto Vezzani

s.pini@unimore.it, guido.borghi@unimore.it, roberto.vezzani@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

softech-ict Centro Interdipartimentale di Ricerca Softech: ICT per le Imprese