Thomas Guionnet
Marc Rivière
Mickaël Raulet

Mile-High Video (MHV) 2024

AI-BASED MONOCULAR DEPTH MAP ESTIMATION APPLIED TO A VIDEO ENCODING PIPELINE

Captivate your audience

DEPTH PERCEPTION

Binocular vision

- Human vision
- Depth from eyes convergence

DEPTH PERCEPTION

Monocular vision

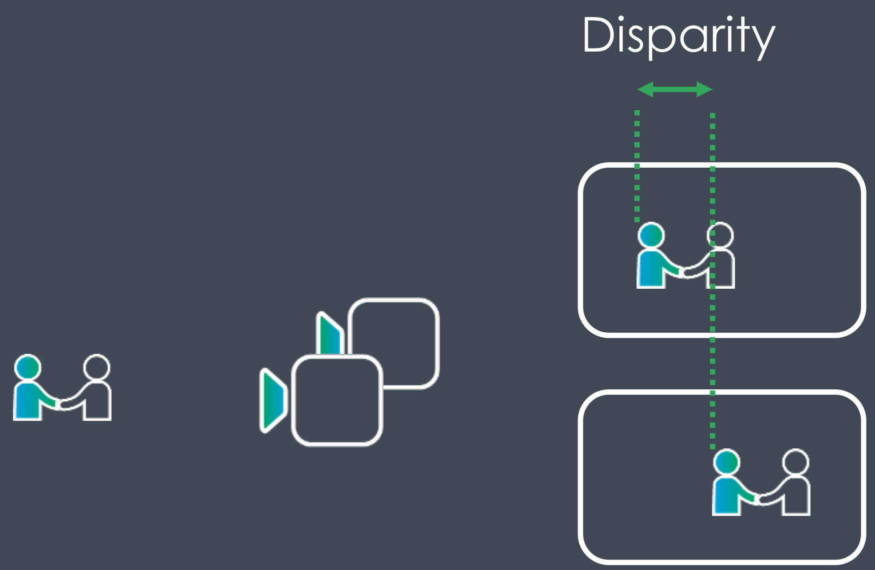
- Human cognitive process
- Depth from scene understanding and structure
- Instantaneous

DEPTH MAPS

- Pixel level depth information
  - Absolute
  - Relative

DEPTH ESTIMATION

- Binocular
  - Stereo capture
  - Depth from disparity



- Monocular
  - Single image

Deep learning

OUTLINE

1. Monocular depth estimation
2. Video coding pipeline
3. Frame interpolation
4. Perspectives

Monocular depth estimation

STATE OF THE ART

Best performance: AdaBins

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. AdaBins: Depth Estimation Using Adaptive Bins. IEEE Computer Society, 4008–4017.

Best speed/performance: PyD-Net

Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. 2018.
Towards real-time unsupervised monocular depth estimation on CPU.
In 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). 5848–5854.

Papier	Abs-Rel ↓	Training Dataset
(Bhat, Alhashim et Wonka, 2021)	**0,058**	KITTI
(Fu et al., 2018)	0,072	KITTI
(Yin, Liu et Shen, 2021)	0,072	KITTI
(Guo et al., 2018)	0,096	KITTI
(Cho et al., 2021)	0,095	KITTI + Cityscapes + DIML/CVL
(Alhashim et Wonka, 2019)	0,093	KITTI
(Luo et al., 2018)	0,094	KITTI
(Guizilini et al., 2020)	0,104	CityScapes -> KITTI
(Shu et al., 2020)	0,104	KITTI
(Lyu et al., 2020)	0,104	CityScapes -> KITTI
(Xu et al., 2018)	0,122	KITTI
(Godard et al., 2019)	0,115	KITTI Stereo
(Atapour-Abarghouei et Breckon, 2018	0,110	KITTI
(Godard et al., 2017)	0,114	KITTI
(Garg et al., 2016)	0,169	KITTI
(Li et al., 2020)	0,130	KITTI
(Bian et al., 2019)	0,128	KITTI + CityScapes
(Li et Snavely, 2018)	0,139	MegaDepth -> KITTI
(Ranftl et al., 2020)	0,157	Mix
(Poggi et al., 2018)	0,146	CityScapes -> KITTI
(Aleotti et al., 2020)	0,162	WILD
(Baig et Torresani, 2016)	0,206	KITTI
(Zhou et al., 2017)	0,198	CityScapes -> KITTI
(Liu et al., 2016)	0,217	KITTI
(Eigen, Puhrsch et Fergus, 2014)	0,190	KITTI
(Li et al, 2019)	0,227	
(Wang et al., 2019)	0,230	KITTI

Absolute Relative Difference

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\hat{y}_i}$$

KITTI dataset (2012)

- Autonomous vehicles
- 93k images
- Lidar generated ground truth

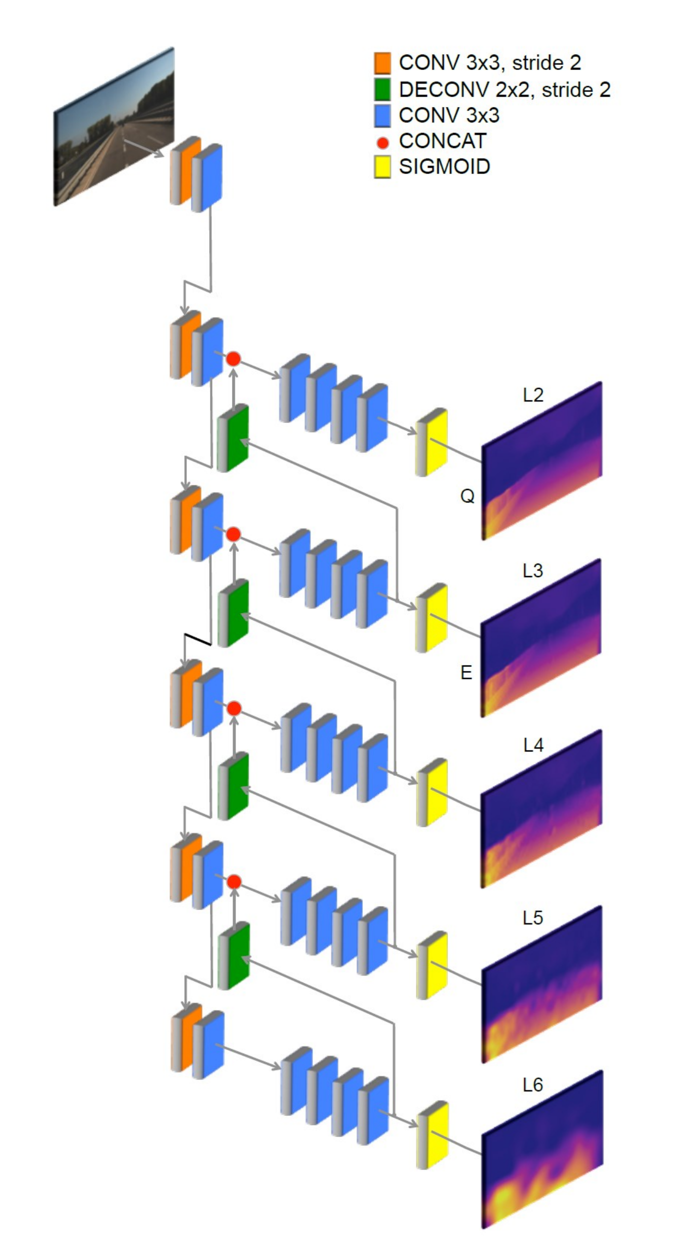
OPTIMIZATION

Resolution	480x288		1920x1080	
Platform	**GPU**	**CPU**	**GPU**	**CPU**
AdaBins	1.4 fps	2 fps	< 0.1 fps	< 0.1 fps
Pyd-Net	27 fps	38 fps	2.6 fps	7.5 fps

GPU 4x NVIDIA Tesla K40m, CPU Intel Xeon Platinium 8268

Resolution	480x288		1920x1080	
Framework	**TensorFlow**	**OpenVino**	**TensorFlow**	**OpenVino**
Pyd-Net L1	26.3 fps	147 fps	3.7 fps	7.7 fps
Pyd-Net L2	32 fps	**416** fps	7.3 fps	**21** fps

CPU AMD Ryzen 9 5900X 12-Core CPU

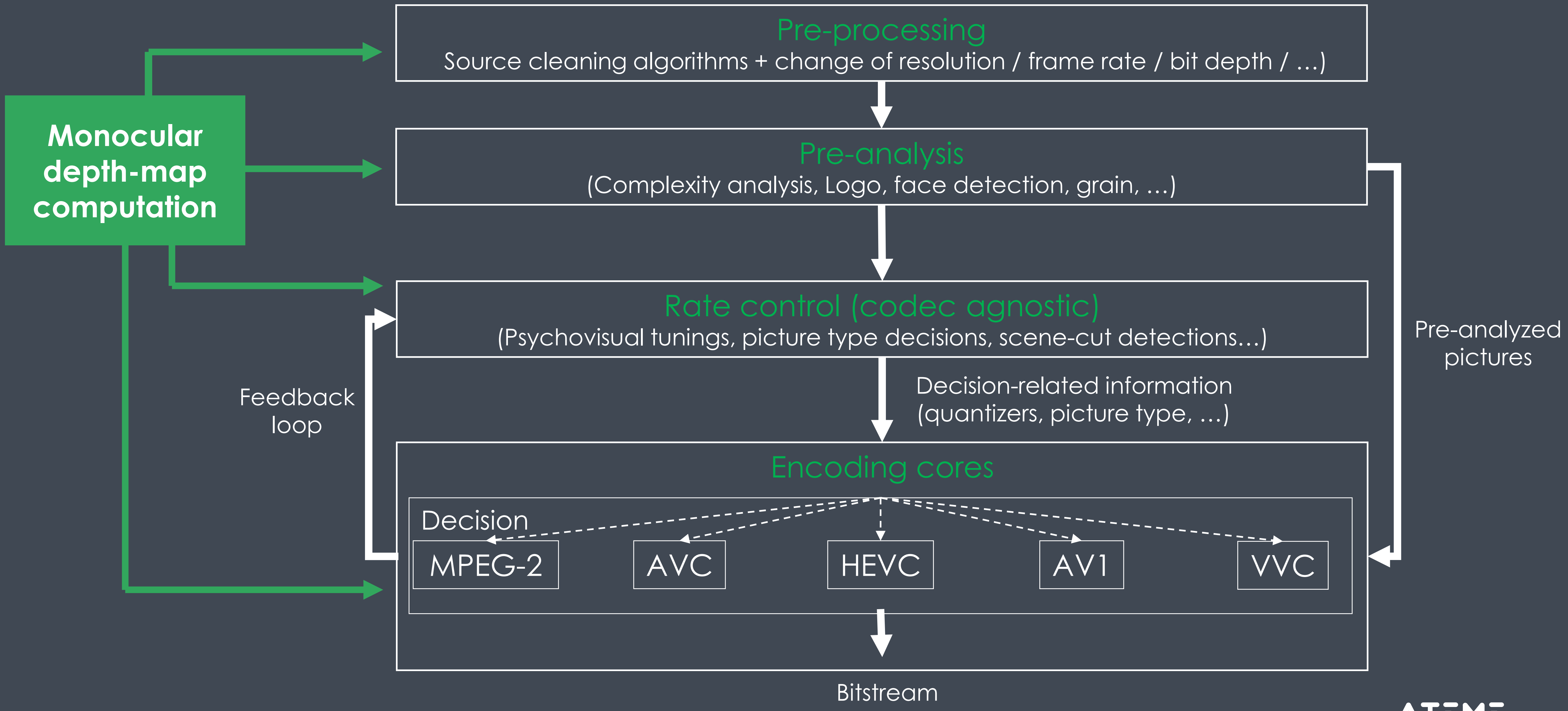


ATEME

2

Video coding pipeline

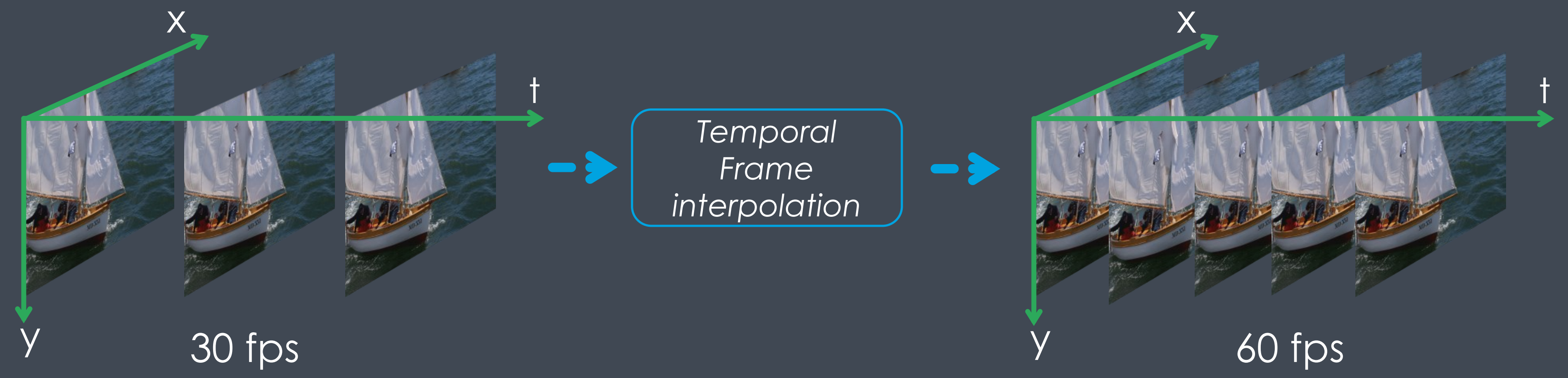
CODING PIPELINE



3

Frame interpolation

INCREASING VIDEO FRAMERATE

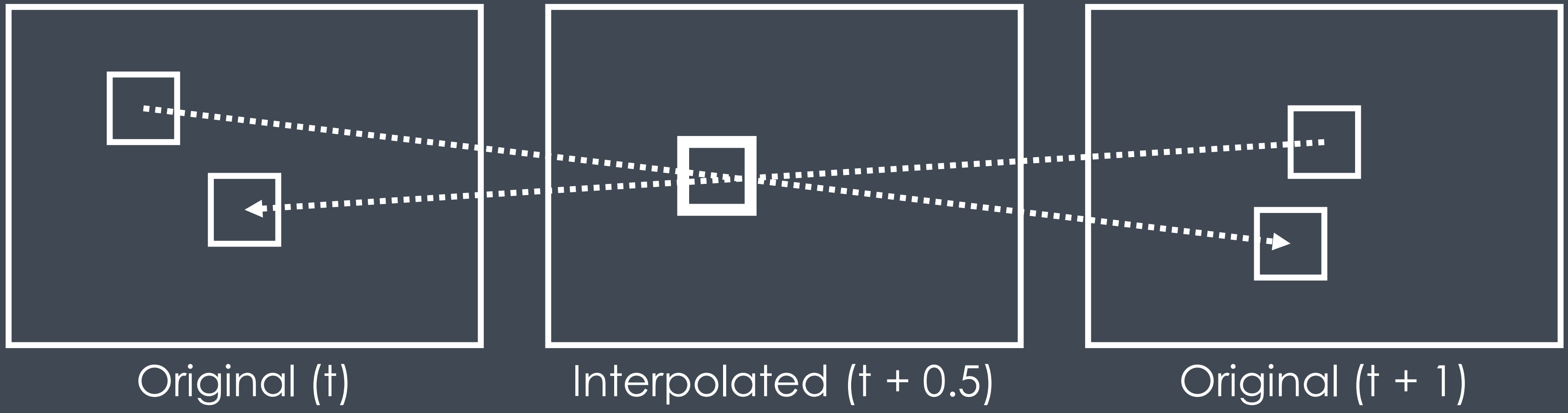


Original (t)

Interpolated (t + 0.5)

Original (t + 1)

MULTI-HYPOTHESIS FRAME INTERPOLATION



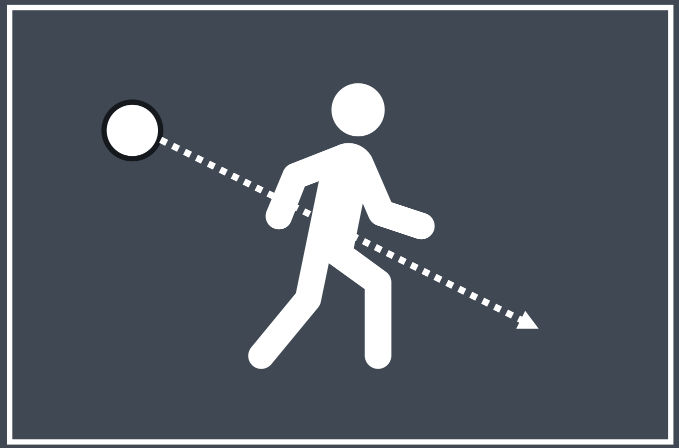
Snježana Rimac-Drlje and Denis Vranješ. 2016. Fast frame-rate upconversion method for video enhancement. In 2016 International Conference on Systems, Signals and Image Processing (IWSSIP). 1–4.

- Forward and backward motion estimation
- Weighted sum of all candidates
- Each pixel of the interpolated frame may receive from zero to N candidates

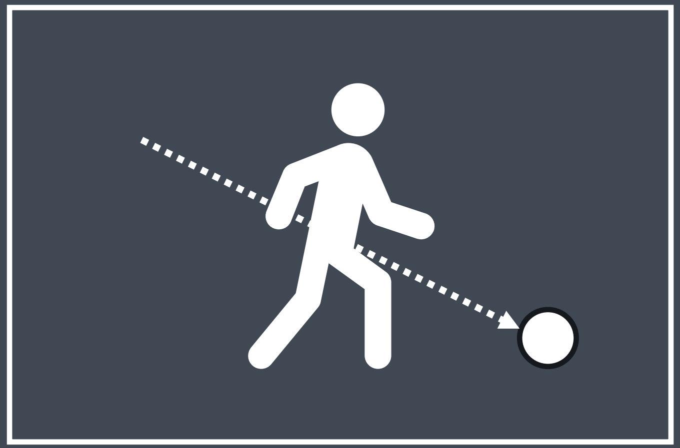
ATEME

NON DECIDABLE FRAME INTERPOLATION

Interpolated t+0.5



Original t



Original t+1

Problem: is the ball behind or in front of the player?

DEPTH BASED FRAME INTERPOLATION

- Converting depth map into a weighting function

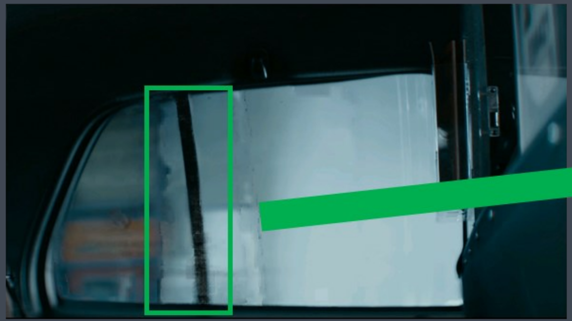
$$W_{D,t}(x,y) = k \cdot \left(1 - \frac{D_t(x,y)}{D_{max,t}}\right)$$

- Weighting pixel candidates for frame interpolation

RESULT EXAMPLE

Interpolated (t + 0.5)

Depth map



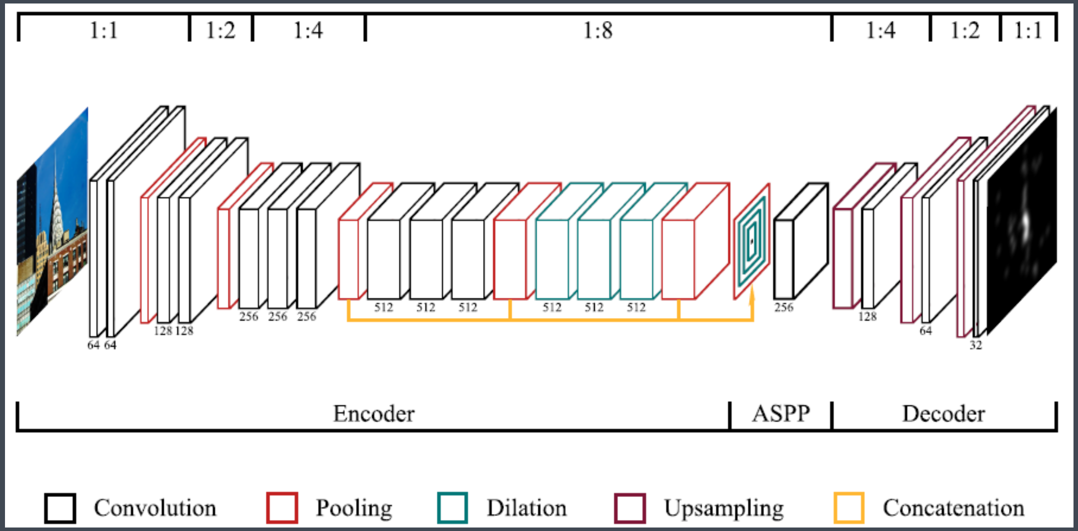
Depth weighted Interpolated (t + 0.5)

4 Perspectives

- Monocular depth map estimation achievable in real-time on regular CPU hardware
- Temporal frame interpolation enhanced thanks to depth information
- Many other possible usages
  - Rate-control
  - Segmentation
- Usefulness highly dependent on depth estimation precision

EXAMPLE: SALIENCY

- Saliency
  - Identifying the regions we are looking at
- Input saliency information to the rate control
  - 17% average bitrate gain [1]
- **Depth is a major information for saliency estimation**
  - Use depth maps as a crude approximation
  - Complement depth maps with a light saliency estimator



Kroner, A. et al. Contextual encoder–decoder network for visual saliency prediction. Neural Networks, 129, 261-270.

[1] Sébastien Pelurson, Josselin Cozanet, Thomas Guionnet, Mohsen Abdoli, and Thibaud Biatek. 2022. AI-Based Saliency-Aware Video Coding. SMPTE Motion Imaging Journal 131, 4 (2022), 21–29.

Captivate your audience

THANK YOU.