Optical Flow and Mode Selection for Learning-based Video Coding

Paper ID: 113

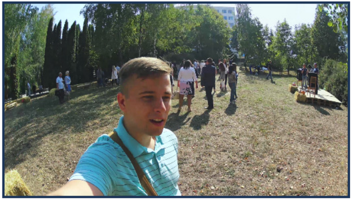
Théo LADUNE[1,2], Pierrick PHILIPPE[1], Wassim HAMIDOUCHE[2], Lu ZHANG[2], Olivier DÉFORGES[2]

[1]Orange Labs, France — [2]INSA Rennes, France
theo.ladune@orange.com

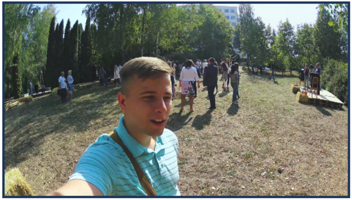
IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Sept. 2020

Introduction

- Successive frames of a video are **highly correlated**



$\mathbf{x}_{t-1}$



$\mathbf{x}_t$

Two successive frames of a video

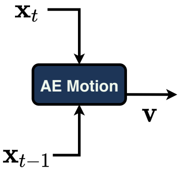
- Video codecs **save rate** by sending $\mathbf{x}_t$ with **inter frame** coding
    1. Computing $\tilde{\mathbf{x}}_t$, a **prediction** of $\mathbf{x}_t$, based on already received frames and **motion information**
    2. Transmitting only the **prediction error**: $r = \mathbf{x}_t - \tilde{\mathbf{x}}_t$

Introduction – Previous work

- Learning-based codecs[1,2,3] implement inter frame coding with

[1]Lu *et al.*, *DVC: an end-to-end deep video compression framework*, **CVPR 19**
[2]Liu *et al.*, *Learned video compression via joint spatial-temporal correlation exploration*
[3]Djelouah *et al.*, *Neural inter-frame compression for video coding*, **ICCV 19**

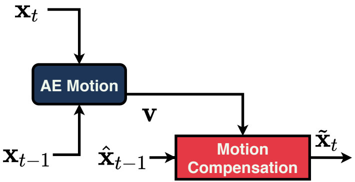
- Learning-based codecs[1,2,3] implement inter frame coding with
    1. One Auto-Encoder (AE) to compute and convey the **optical flow** $\mathbf{v}$



[1] Lu *et al.*, *DVC: an end-to-end deep video compression framework*, **CVPR 19**
[2] Liu *et al.*, *Learned video compression via joint spatial-temporal correlation exploration*
[3] Djelouah *et al.*, *Neural inter-frame compression for video coding*, **ICCV 19**

Introduction – Previous work

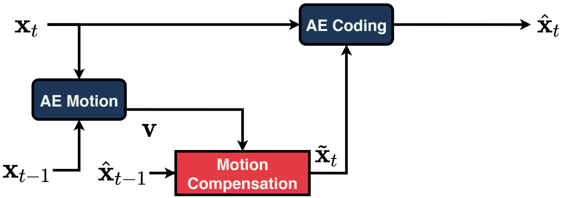
- Learning-based codecs[1,2,3] implement inter frame coding with
  1. One Auto-Encoder (AE) to compute and convey the **optical flow** $\mathbf{v}$
  2. Interpolation of a reference frame to perform **motion compensation**



[1] Lu *et al.*, *DVC: an end-to-end deep video compression framework*, **CVPR 19**
[2] Liu *et al.*, *Learned video compression via joint spatial-temporal correlation exploration*
[3] Djelouah *et al.*, *Neural inter-frame compression for video coding*, **ICCV 19**

Introduction – Previous work

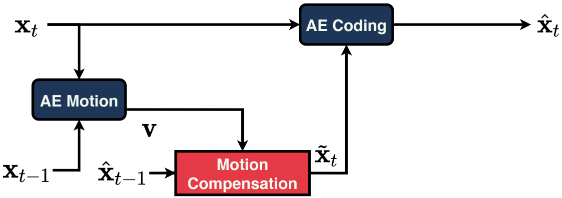
- Learning-based codecs[1,2,3] implement inter frame coding with
  1. One Auto-Encoder (AE) to compute and convey the **optical flow** $\mathbf{v}$
  2. Interpolation of a reference frame to perform **motion compensation**
  3. One AE to perform **residual** (prediction error) coding



[1] Lu *et al.*, *DVC: an end-to-end deep video compression framework*, **CVPR 19**
[2] Liu *et al.*, *Learned video compression via joint spatial-temporal correlation exploration*
[3] Djelouah *et al.*, *Neural inter-frame compression for video coding*, **ICCV 19**

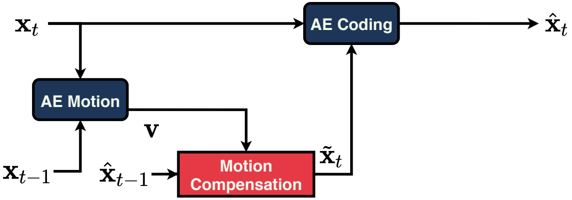
Introduction – Previous work

- Learning-based codecs[1,2,3] implement inter frame coding with
  1. One Auto-Encoder (AE) to compute and convey the **optical flow** $\mathbf{v}$
  2. Interpolation of a reference frame to perform **motion compensation**
  3. One AE to perform **residual** (prediction error) coding



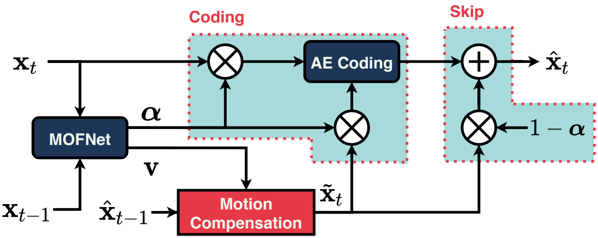
- Drawbacks
  - **Separate training** of both AEs with proxy metrics
  - **No mode competition** unlike classical codecs (intra/inter/skip)
  - **Residual coding** not the best mixture of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$

Introduction – Contributions



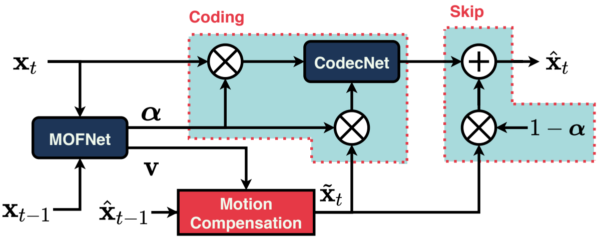
Drawback	*Contributions*
No mode competition	
Residual coding	
Separate training	

Introduction – Contributions



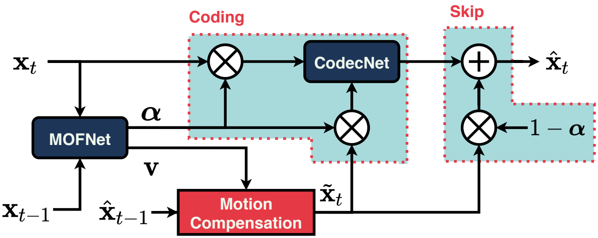
Drawback	*Contributions*
No mode competition	MOFNet: Pixel-wise **mode competition** Coding vs. Skip ($\tilde{\mathbf{x}}_t$ copy)
Residual coding	
Separate training	

Introduction – Contributions



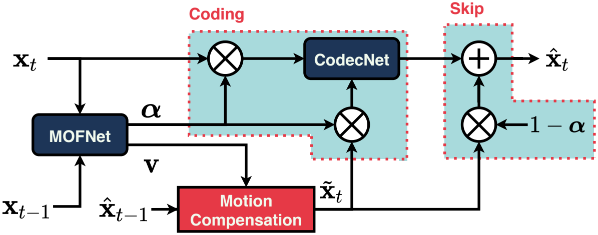
Drawback	*Contributions*
No mode competition	MOFNet: Pixel-wise **mode competition** Coding vs. Skip ($\tilde{\mathbf{x}}_t$ copy)
Residual coding	CodecNet: **arbitrary mixture** of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$
Separate training	

Introduction – Contributions



Drawback	*Contributions*
No mode competition	MOFNet: Pixel-wise **mode competition** Coding vs. Skip ($\tilde{\mathbf{x}}_t$ copy)
Residual coding	CodecNet: **arbitrary mixture** of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$
Separate training	Skip mode fosters optical flow learning $\rightarrow$ enable **end-to-end training**

Introduction – Contributions



Drawback	*Contributions*
No mode competition	MOFNet: Pixel-wise **mode competition** Coding vs. Skip ($\tilde{\mathbf{x}}_t$ copy)
Residual coding	CodecNet: **arbitrary mixture** of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$
Separate training	Skip mode fosters optical flow learning $\rightarrow$ enable **end-to-end training**

- Perform on par with HEVC on a P-frame coding task

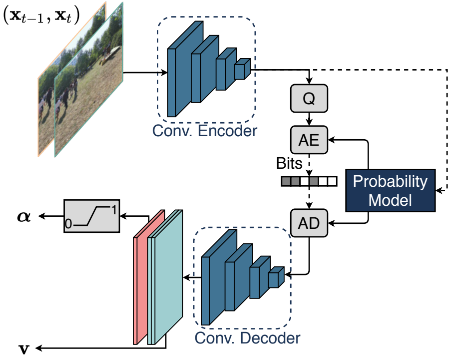
Proposed system – MOFNet

- MOFNet architecture: **standard AE** with hyperprior (AE-HP)[1]

[1] Minnen *et al.*, *Joint Autoregressive and Hierarchical Priors for Learned Image Compression*, **NIPS 18**

Proposed system – MOFNet

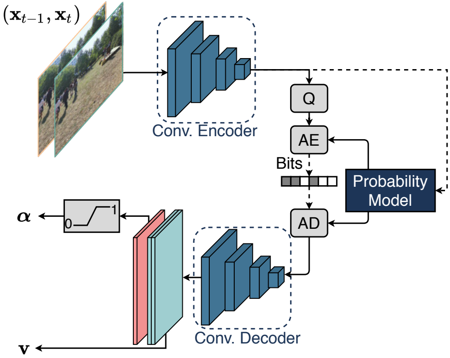
- MOFNet architecture: **standard AE** with hyperprior (AE-HP)[1]



[1] Minnen *et al.*, *Joint Autoregressive and Hierarchical Priors for Learned Image Compression*, **NIPS 18**

Proposed system – MOFNet

- MOFNet architecture: **standard AE** with hyperprior (AE-HP)[1]

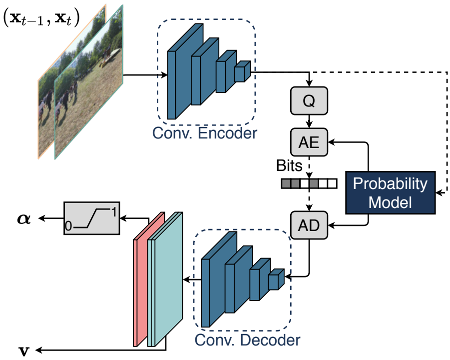


- Prediction: $\tilde{\mathbf{x}}_t = w(\hat{\mathbf{x}}_{t-1}, \mathbf{v})$, with $\begin{cases} \mathbf{v} \in \mathbb{R}^{2\times H\times W} \text{ the **optical flow**,} \\ w \text{ a bilinear warping} \end{cases}$

[1]Minnen *et al.*, *Joint Autoregressive and Hierarchical Priors for Learned Image Compression*, **NIPS 18**

Proposed system – MOFNet

- MOFNet architecture: **standard AE** with hyperprior (AE-HP)[1]

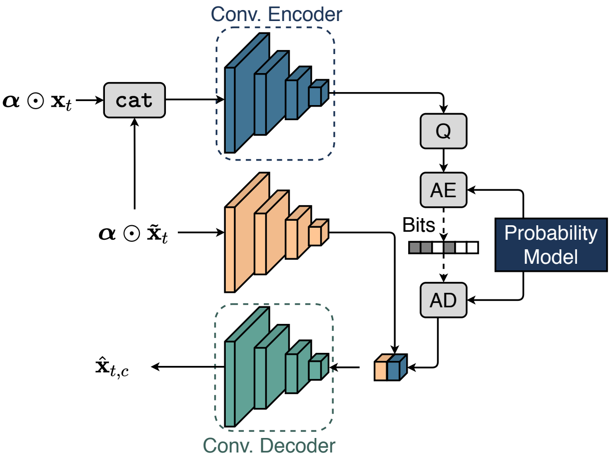


- Prediction: $\tilde{\mathbf{x}}_t = w(\hat{\mathbf{x}}_{t-1}, \mathbf{v})$, with $\begin{cases} \mathbf{v} \in \mathbb{R}^{2\times H \times W} \text{ the } \textbf{optical flow}, \\ w \text{ a bilinear warping} \end{cases}$
- $\boldsymbol{\alpha} \in [0,1]^{H\times W}$ arbitrates **2 coding modes** $\begin{cases} \text{Transmission by CodecNet} \\ \text{Skip mode } (\textit{Prediction copy}) \end{cases}$

[1] Minnen *et al.*, *Joint Autoregressive and Hierarchical Priors for Learned Image Compression*, **NIPS 18**

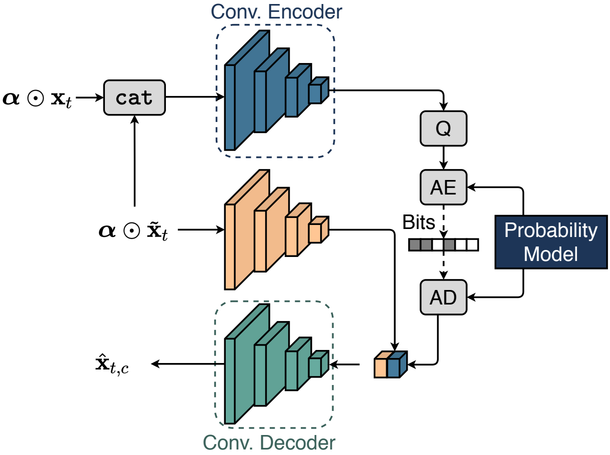
Proposed system – CodecNet implementation

- CodecNet is an AE-HP which conveys the **areas selected** by α



Proposed system – CodecNet implementation

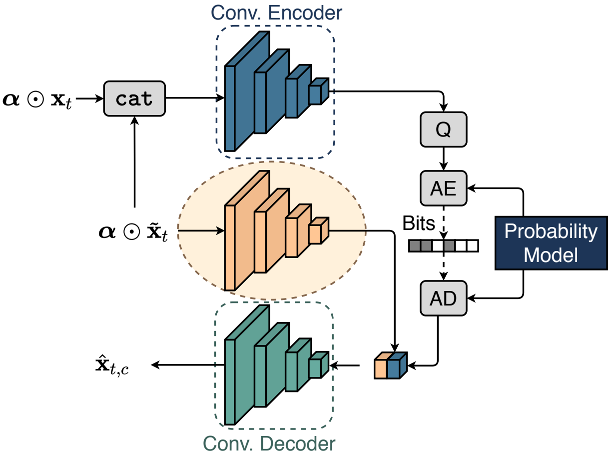
- CodecNet is an AE-HP which conveys the **areas selected** by $\boldsymbol{\alpha}$



- Perform **conditional coding**: complex non-linear mixture of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$

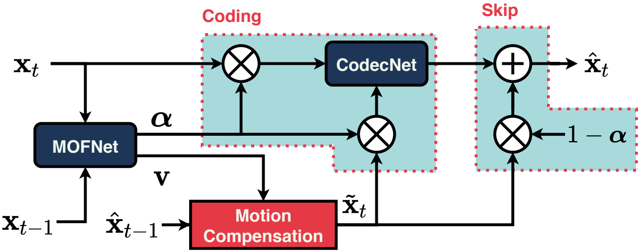
Proposed system – CodecNet implementation

- CodecNet is an AE-HP which conveys the **areas selected** by $\boldsymbol{\alpha}$



- Perform **conditional coding**: complex non-linear mixture of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$
- Supplementary transform to extract **useful features** from $\tilde{\mathbf{x}}_t$

Proposed system

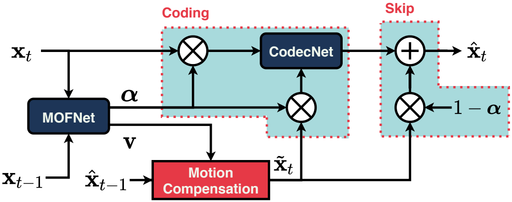


- MOFNet & CodecNet are combined to obtain the complete system

- Two competing coding modes
  - α close to 0 corresponds to **Skip mode**
  - α close to 1 corresponds to **Conditional coding**

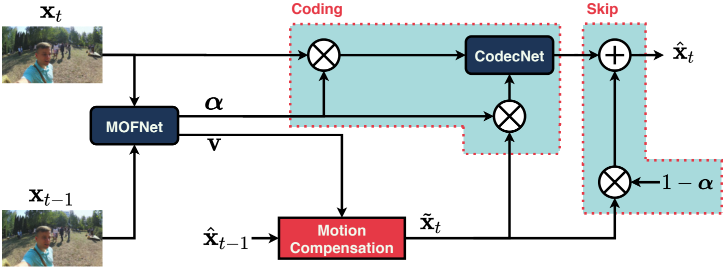
- **End-to-end** training with rate distortion cost
  - **No dedicated** loss term for α or $\mathbf{v}$: $\mathcal{L}_\lambda = \mathrm{D}(\mathbf{x}_t, \hat{\mathbf{x}}_t) + \lambda\,(R_m + R_c)$

- CLIC20 P-frame test conditions[1]
  - Quality metric is **MS-SSIM** *i.e.* $\mathrm{D}(\mathbf{x}_t, \hat{\mathbf{x}}_t) = 1 - \text{MS-SSIM}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$
  - Reference frame is assumed **lossless**: $\hat{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1}$



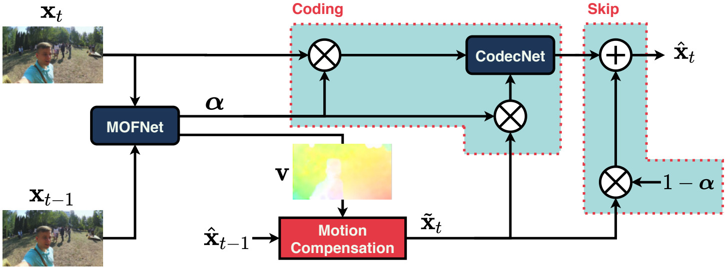
[1] *Challenge on Learned Image Compression*, `www.compression.cc`, **CVPR 20**

Visualisation



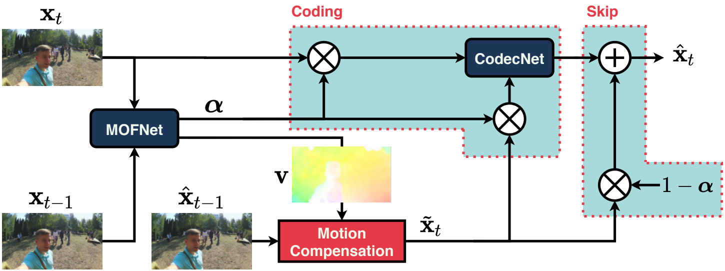
Input frames $(\mathbf{x}_{t-1}, \mathbf{x}_t)$

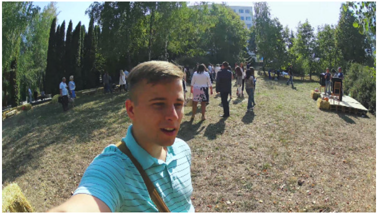
Visualisation



Optical flow $\mathbf{v}$

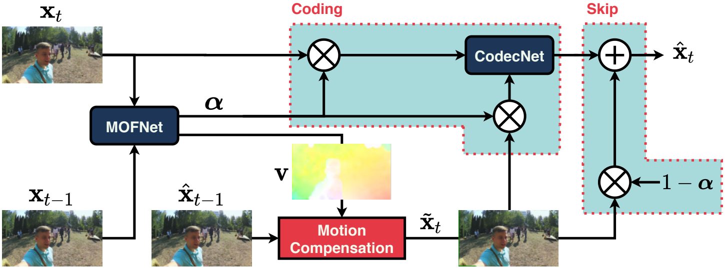
Visualisation

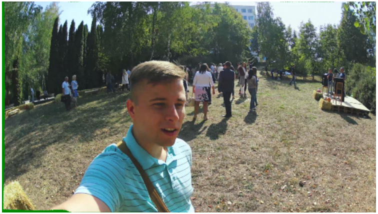




Reference frame $\hat{\mathbf{x}}_{t-1}$

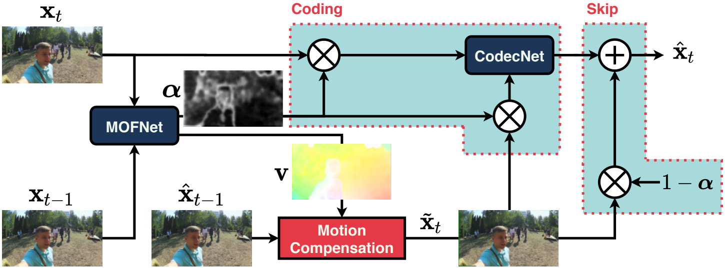
Visualisation





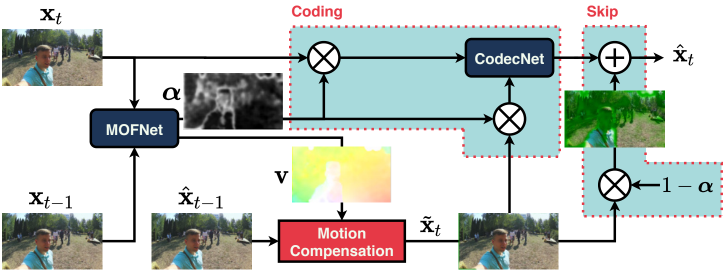
$$\textit{Prediction } \tilde{\mathbf{x}}_t = w(\hat{\mathbf{x}}_{t-1}, \mathbf{v})$$

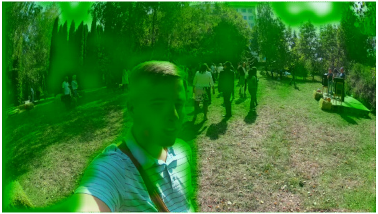
Visualisation



Coding mode selection α

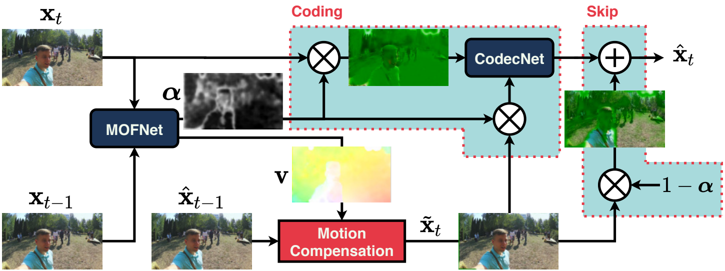
Visualisation





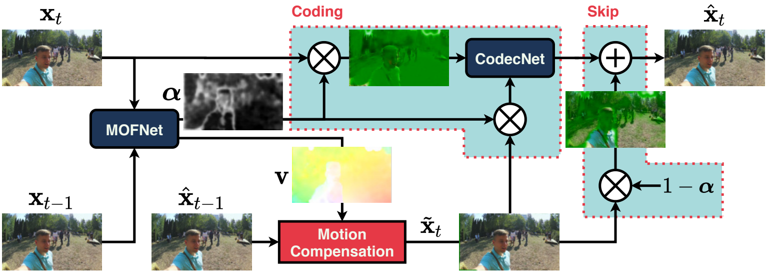
Skip part $(1-\boldsymbol{\alpha}) \odot \tilde{\mathbf{x}}_t$

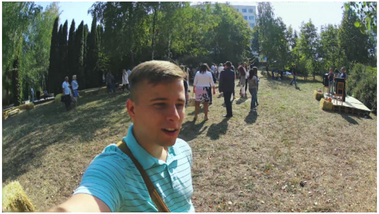
Visualisation



Coding part $\boldsymbol{\alpha} \odot \mathbf{x}_t$

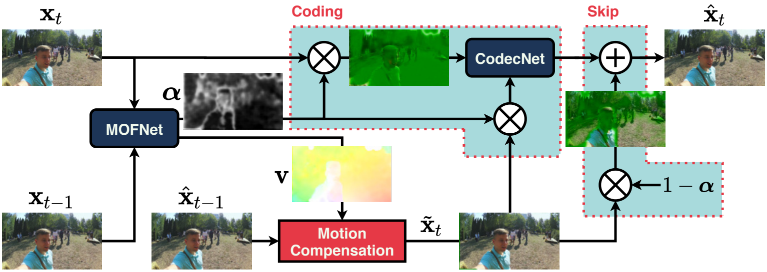
Visualisation

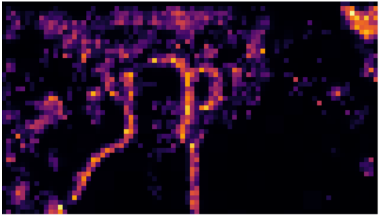




$$\textit{System output } \hat{\mathbf{x}}_t = c(\alpha \odot \mathbf{x}_t, \alpha \odot \tilde{\mathbf{x}}_t) + (1-\alpha) \odot \tilde{\mathbf{x}}_t$$

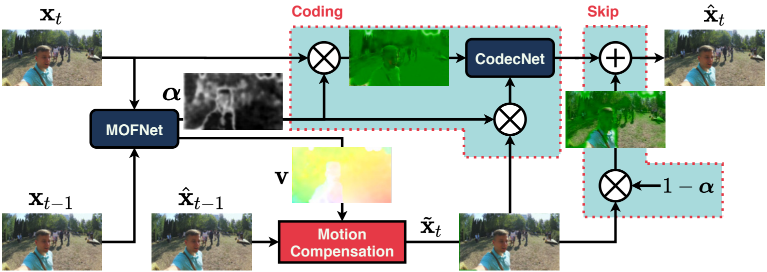
Visualisation

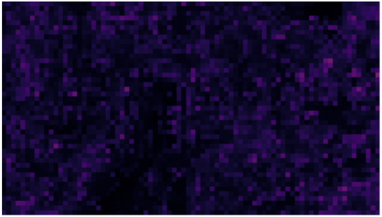




CodecNet rate $\mathrm{R}_c = 0.022$ *bpp*

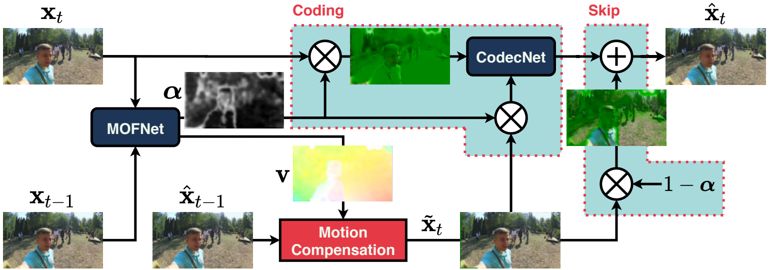
Visualisation





MOFNet rate $\mathrm{R}_m = 0.019$ *bpp*

Visualisation



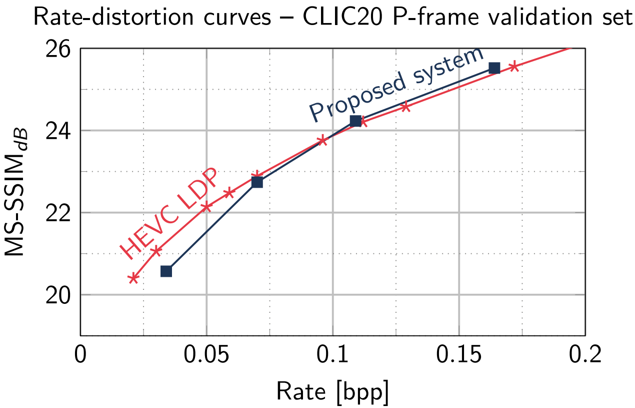
- MOFNet learns a coding mode selection α and an optical flow $\mathbf{v}$
- The training is driven by a **simple rate-distortion loss**
  - Skip mode presence is an incentive to learn a relevant optical flow

Results – Experimental conditions

- This work follows the CLIC20 P-frame coding test conditions
  - Quality metric is **MS-SSIM**
  - Rate target is around **0.075 bpp**
  - CLIC20 P-frame validation set

- Two experiments carried out
  1. Proposed system vs. **HEVC low-delay P**
  2. **Ablation** study to assess the benefits of skip mode and conditional coding → **Results available in the paper**

Results – Proposed system vs. HEVC



- **Proposed system** performs on par with **HEVC low-delay P**
- It proves the possibility of learning a relevant optical through a mere **RD-cost**

Conclusion

- This paper introduces a **new end-to-end** inter-frame coding scheme based on 2 Auto-Encoders
  - MOFNet: transmit an **optical flow** and a **coding mode selection**
  - CodecNet: perform **conditional coding** of a frame given its prediction
- The proposed coding scheme implements competition between CodecNet and Skip mode
  - Enable optical flow learning **without dedicated loss**
  - Improve performance, making it **competitive with HEVC**
- Future work: adapt the system to handle **several reference** frames to achieve better coding efficiency