IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Motion Magnification

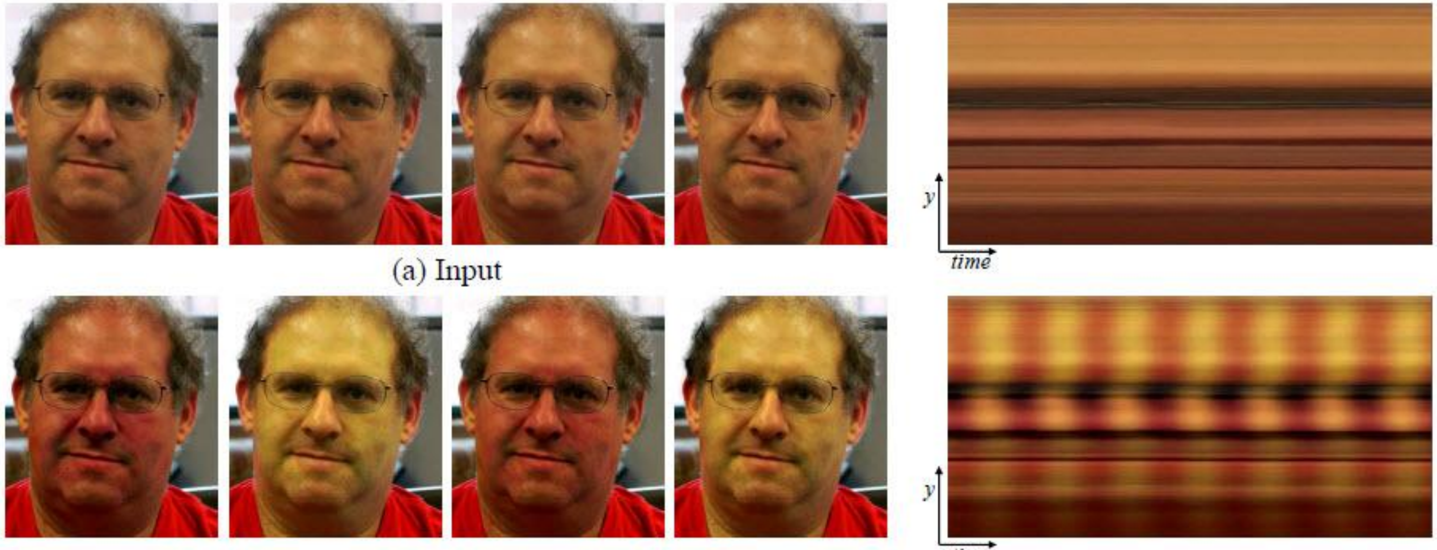
SUMMARY

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Introduction to Motion Magnification

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Discerning small motions => important applications (understanding a building's structural health, measuring a person's vital sign)



(a) Input

(b) Magnified

(c) Spatiotemporal *YT* slices

Example: visualizing the human pulse, by amplifying the periodic color variation

Example: effect of the wind on a crane

Difficulty: How can we distinguish between small motion and noise ?

=> Previous video motion techniques suffer from **noisy outputs** and **excessive blurring**

Their **principle**: decomposing video frames into representations magnifying motion, using hand-designed filters

> Can we use better filters than hand-designed ones ?

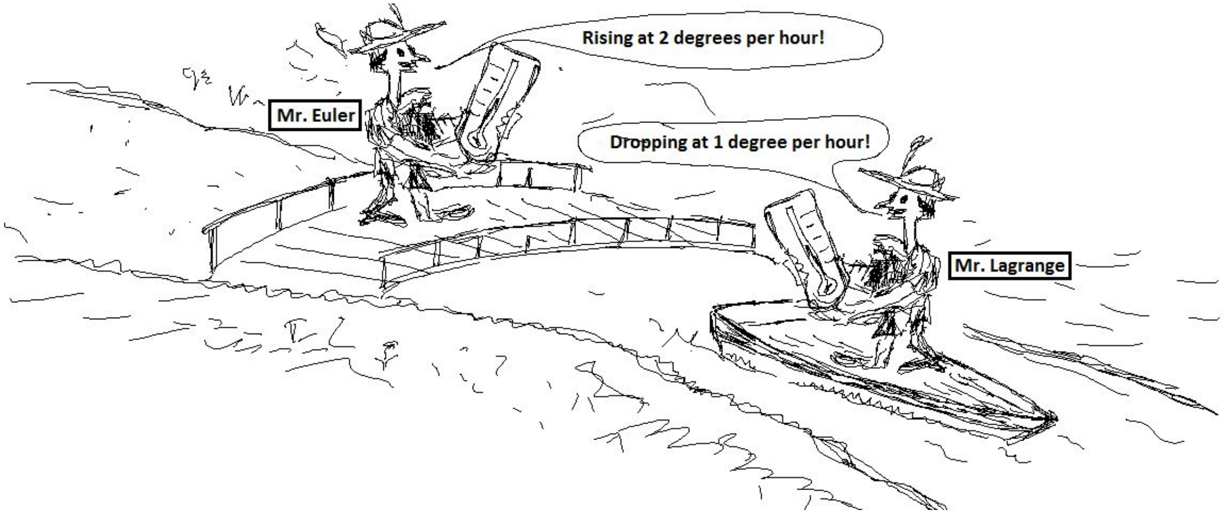
Goal of the paper: learning decomposition filters using deep convolutional neural networks (CNN)

Dataset: a synthetic one simulating small motion

→ Design of a **network** made up of the spatial decomposition filters, the representation manipulator, and the reconstruction filters
Training: two-frames input, magnified difference as the target

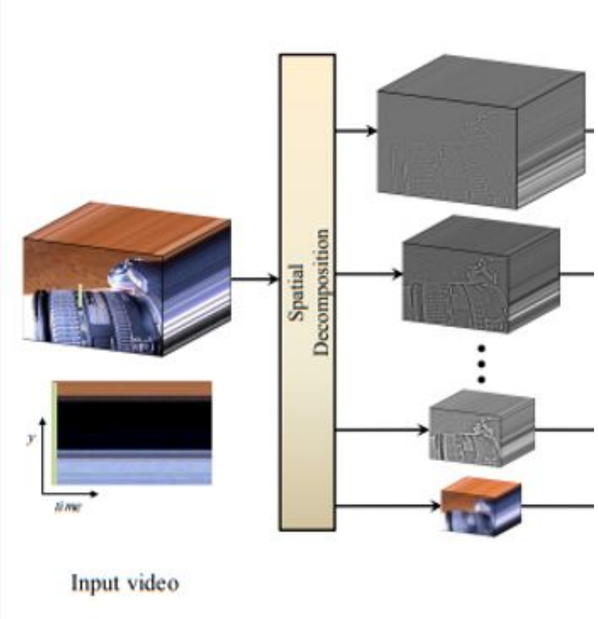
Contributions: high-quality magnification, learned filters generalizing well in real videos

The Eulerian approach vs the Lagrangian approach :



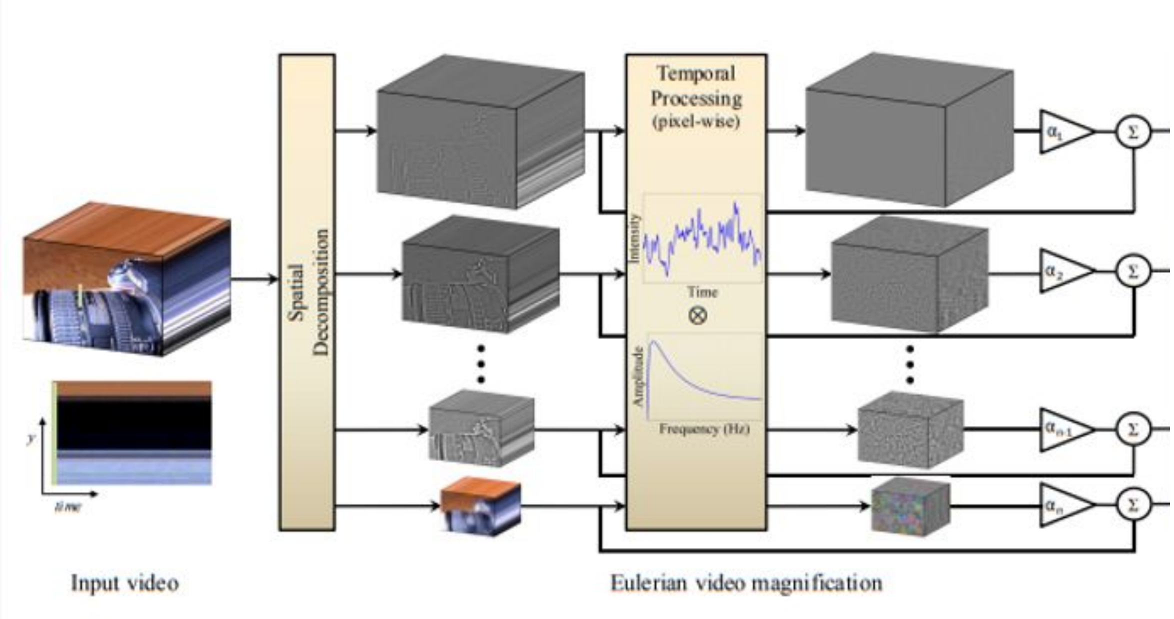
Lagrangian method : extraction the motion field
Eulerian method : representations to manipulate motions without needing to track the motion through space and time

The pipeline :



Filtering to get multiple spatial scales or frequency bands corresponding to different levels of detail

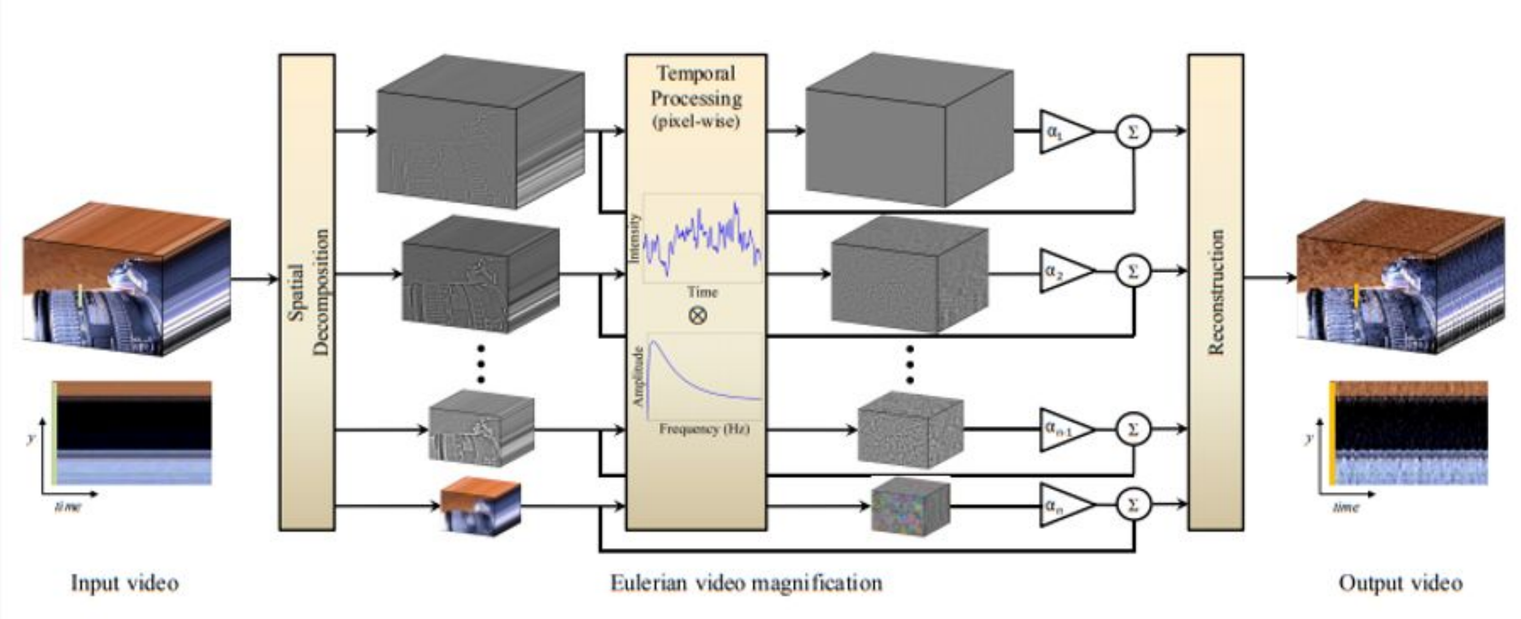
The pipeline :



Temporal filtering : isolate specific motions

Amplification of the motion

The pipeline :



Other methods :

- hand-designed filters like steerable filters for spatial decomposition but noise produced
- temporal filters to isolate motion of interest and to not amplify the noise

Our method :

- decomposition filters are learned from examples with deep convolutional networks (much less noise)

Original images : **Magnified image :**

$$I(x, t0) = f(x) \quad \text{and } I(x, t1) = f(x+\delta(t1)) \;\rightarrow\; I'(x, t1) = f(x + (1 + \alpha)\delta(t))$$

- $\delta(t)$ represents the motion field : we suppose a global translation over time
- α is the magnification factor

Magnification for a motion of interest :

$$\delta'(x, t) = T\,(\delta(x, t))$$

$T\,(\cdot)$ is a temporal bandpass filter

Phase-Based Motion Magnification

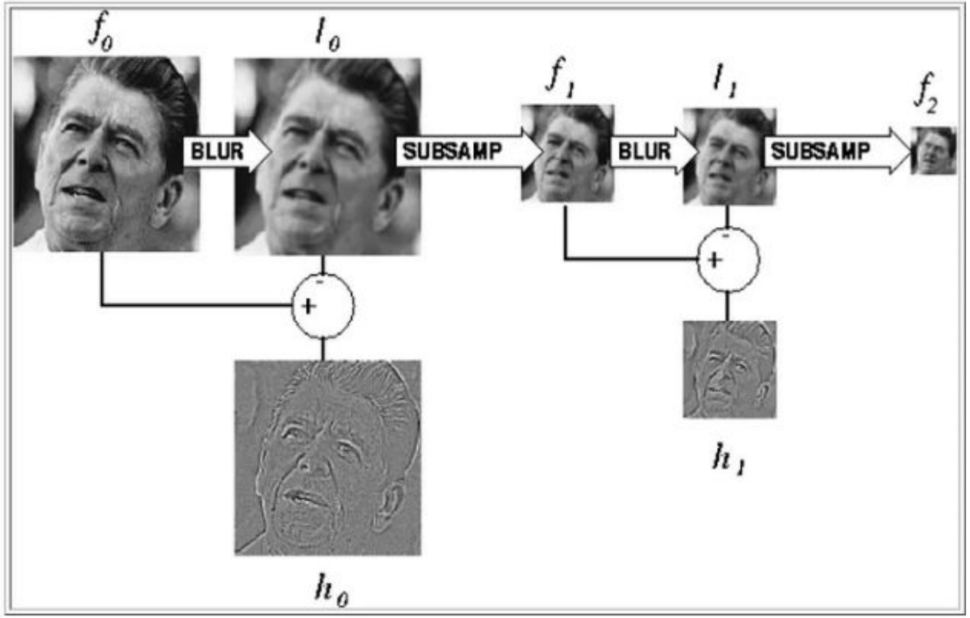
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

An **Image Pyramid** is a multi-scale decomposition of an image.

- Enhanced Information Retrieval
- Scale-Sensitive Analysis
- Improved Feature Detection

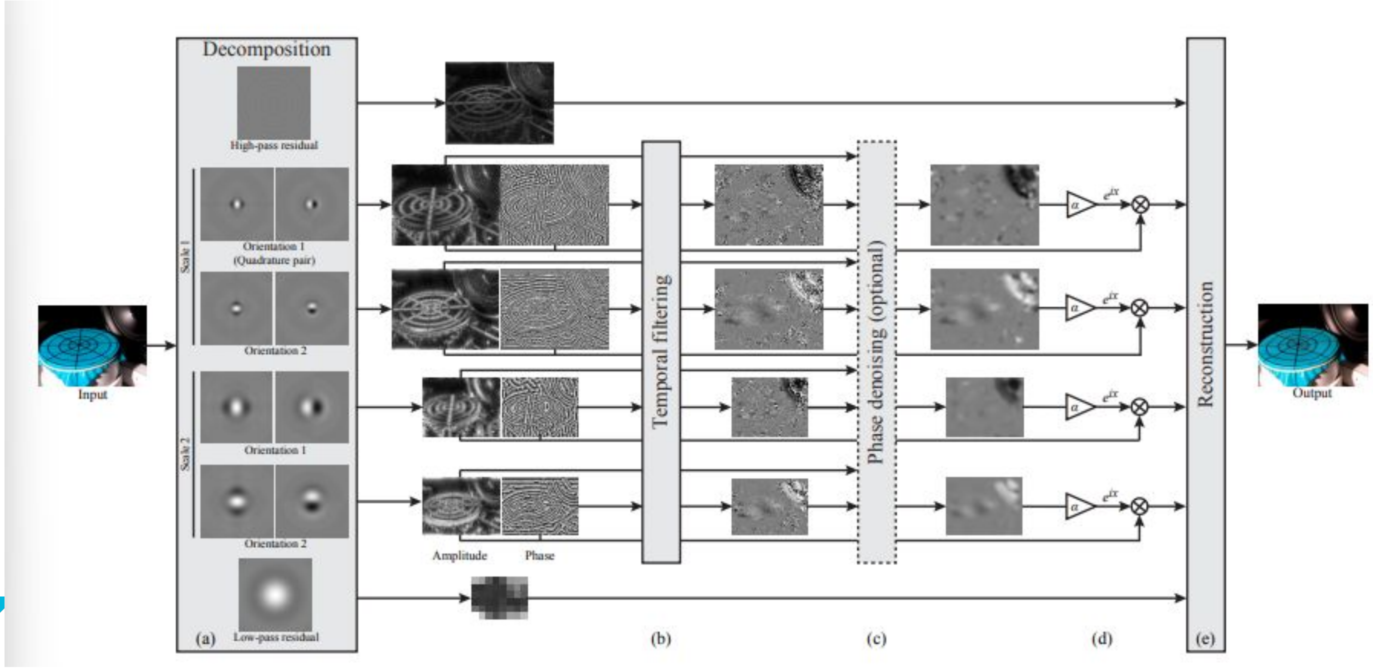
A common example is the **Laplacian Pyramid.**

- Levels consist of residuals between the original and Gaussian blurred images.
- Residuals (h_0, h_1) act as band-pass filtered images.
- Lower scales correspond to lower frequencies.



Example of the Laplacian pyramid (https://medium.com/@itberrios6/steerable-pyramids-6bfd4d23c1)

A Steerable Pyramid is a collection of **filters** at different **sub-bands** which also include separate Low and High Pass components.

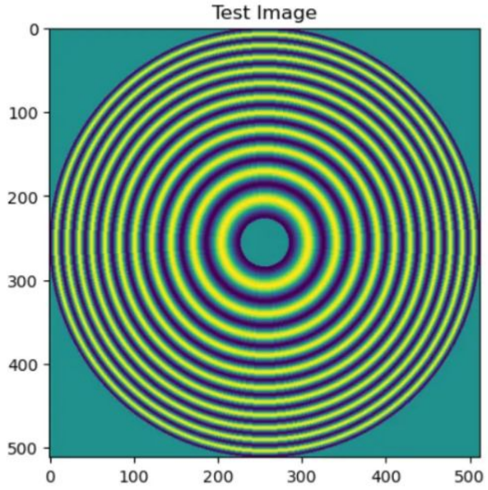


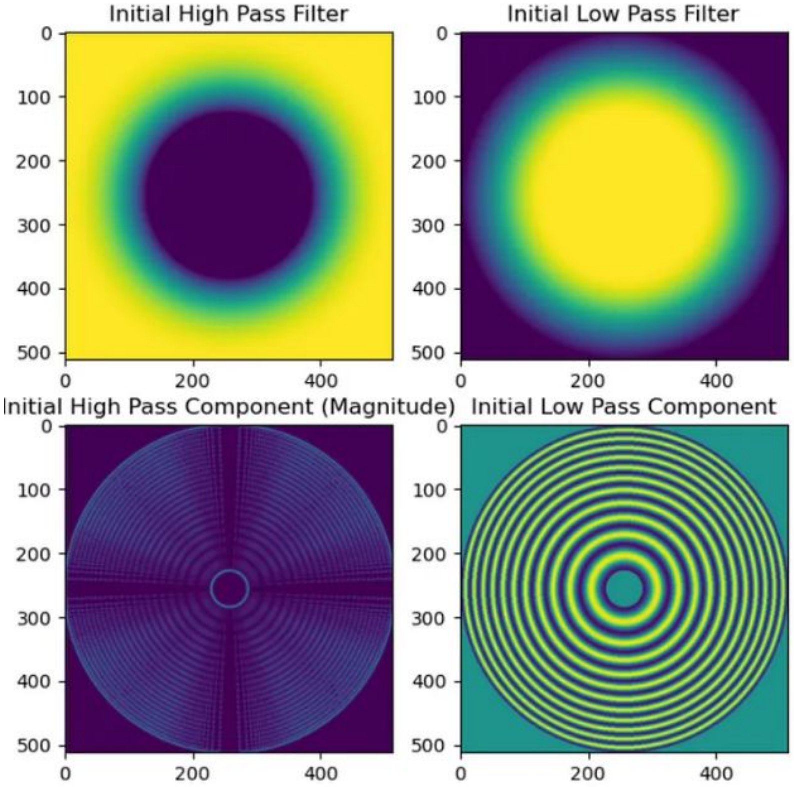
2. Phase-based Motion Magnification

2.1 Complex steerable pyramids

Performing the Decomposition

- **Getting the first High and Low Pass Filters**





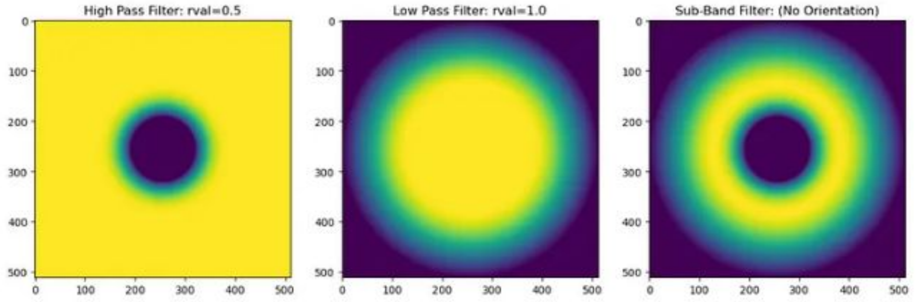
2. Phase-based Motion Magnification
2.1 Complex steerable pyramids

- **Getting the Sub-Band Filters**
  - Filters in a Steerable Pyramid's sub-bands must be **polar separable**.

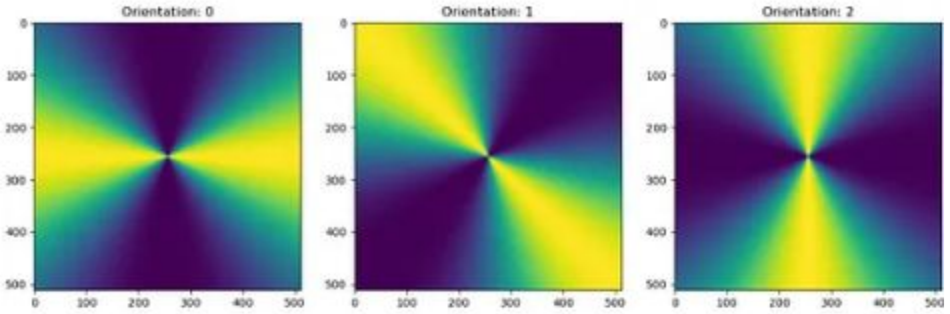
⇒ This entails the ability to decompose them into radial and angular components.

$$B_i(\vec{\omega}) = A(\theta - \theta_i)B(\omega)$$

A : the angular component w.r.t θ
θ_i : the steering/orientation angle
B : the radial component
ω : the frequency band



The construction of the sub-band filter B(w)

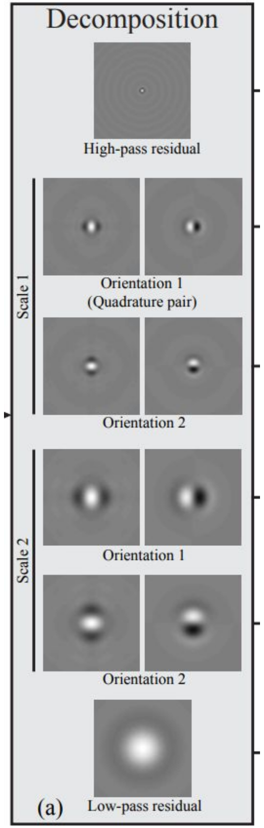


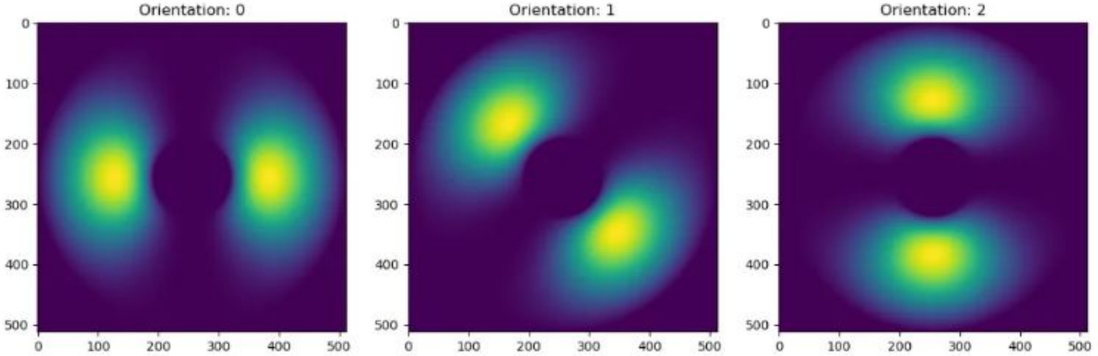
The orientation filters (steered angles): A

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

2. Phase-based Motion Magnification

2.1 Complex steerable pyramids





- Transfer functions(the impulsion response of the filters) are applied to the discrete Fourier transform Î of an image I to create a steerable pyramid.
- This process decomposes the image into different spatial frequency bands $S_{\omega,\theta}$ with corresponding $\tilde{S}_{\omega,\theta}(x,y) = \tilde{I}\Psi_{\omega,\theta}$

Conclusion:

⇒ Each filter isolates a continuous region in the frequency domain, resulting in a localized impulse response in space.
⇒ The resulting spatial frequency bands are localized in space, scale, and orientation.
⇒ The transfer functions of a complex steerable pyramid contain only positive frequencies.
⇒ This approach allows for consideration of both amplitude and phase in the response.

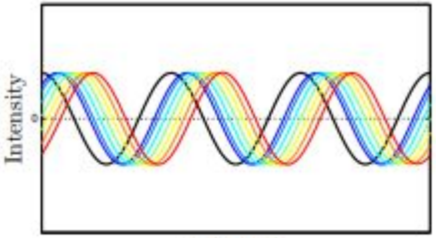
⇒ To magnify the small motion, we modify local phase variations in a complex steerable pyramid representation of the video.

A 1D image intensity profile f under global translation over time is represented as $f(x + \delta(t))$, where $\delta(t)$ is some displacement function.

Goal: synthesize a sequence with modified motion, $f(x + (1 + \alpha)\delta(t))$, for some magnification factor α.

Hint: using the Fourier decomposition, $f(x+\delta(t)) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x+\delta(t))}$

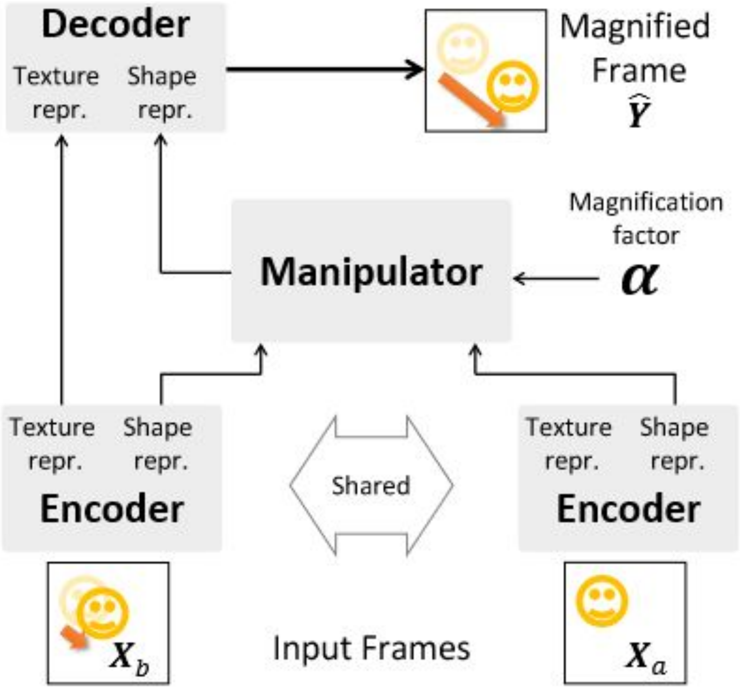
$$\Rightarrow \hat{S}_\omega(x,t) = S_\omega(x,t)e^{i\alpha B_\omega} = A_\omega e^{i\omega(x+(1+\alpha)\delta(t))}$$



(c) Phase-based

Learning-Based Motion Magnification

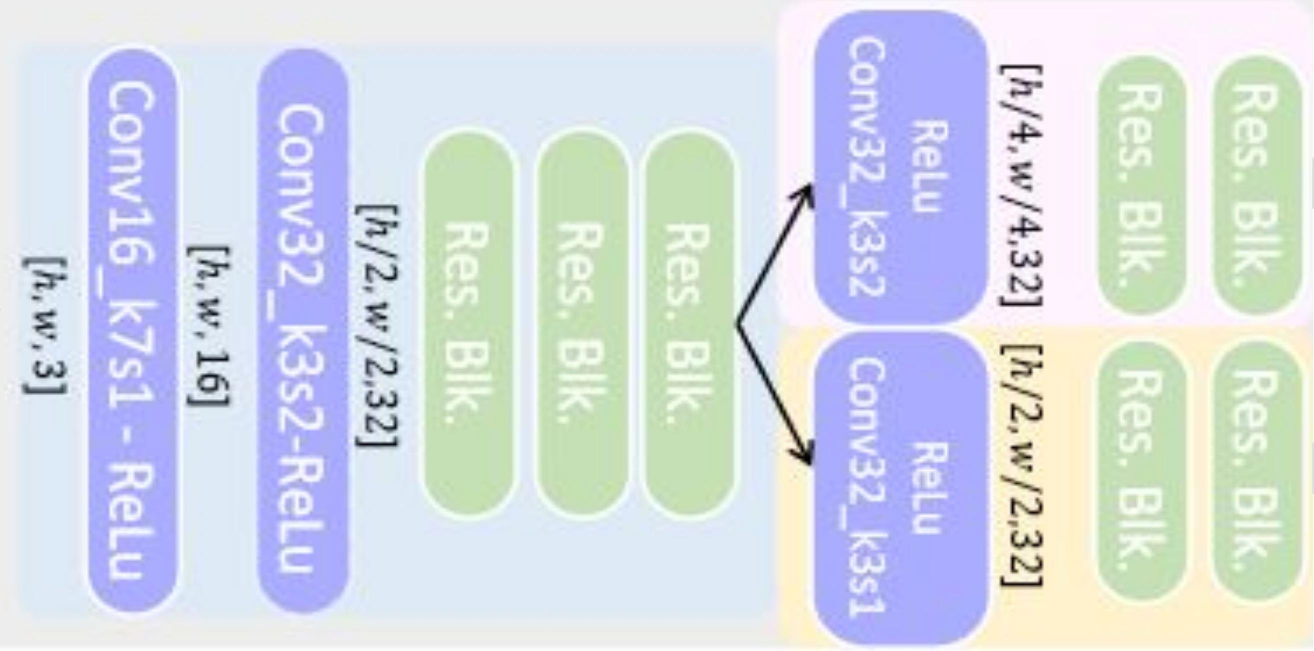
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom



Decoder
Texture repr.
Shape repr.
Magnified Frame
$\hat{Y}$
Magnification factor
α
Manipulator
Texture repr.
Shape repr.
Encoder
Shared
Texture repr.
Shape repr.
Encoder
X_b
Input Frames
X_a

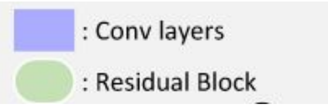
Encoder

Input



Shape Representation (M=Ge,shape(X))

Texture Representation (V=Ge,texture(X))



Why this division ?

Shape Representation (M) → Contains motion-related information

Texture Representation (V) → Contains intensity-related information

Why this division ?

Shape Representation (M) → Contains motion-related information

Texture Representation (V) → Contains intensity-related information

We need to magnify M & Avoid intensity Magnification

Manipulator - 2 Frames

$$G_m(\mathrm{M}_a, \mathrm{M}_b, \alpha) = \mathrm{M}_a + \alpha(\mathrm{M}_b - \mathrm{M}_a).$$

Where
Ma: Shape representation of the first frame A
Mb: Shape representation of the second frame B
α: Magnification factor

Manipulator - 2 Frames

$$G_m(\mathbf{M}_a, \mathbf{M}_b, \alpha) = \mathbf{M}_a + h\left(\alpha \cdot g(\mathbf{M}_b - \mathbf{M}_a)\right)$$

Where

Ma: Shape representation of the first frame A

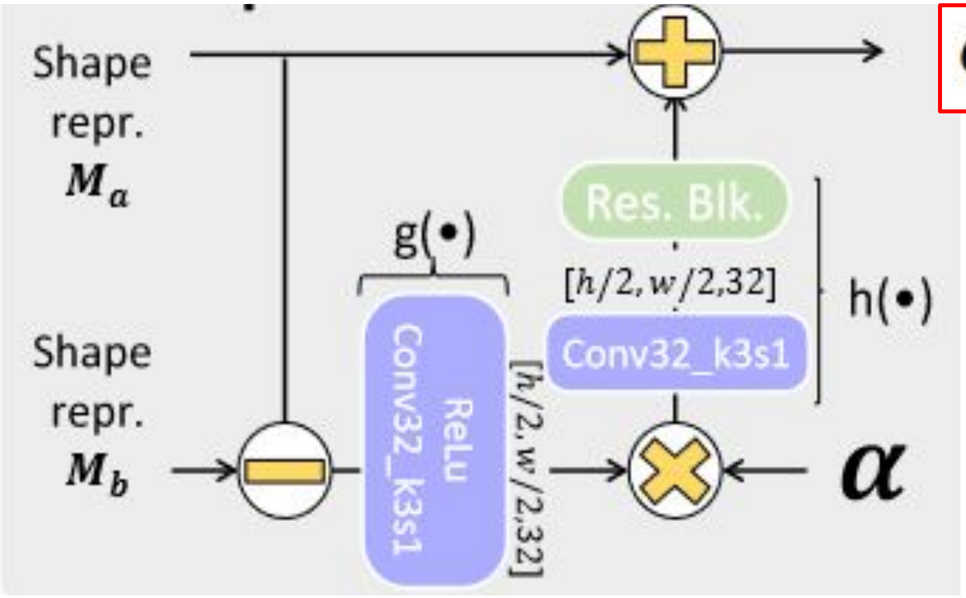
Mb: Shape representation of the second frame B

α: Magnification factor

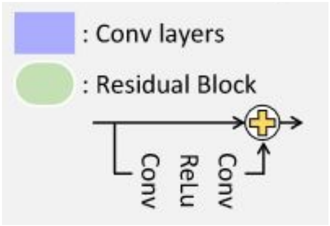
h(.): 3*3 conv + residual block

g(.): 3*3 conv + RELU

Manipulator - 2 Frames



$$G_m(\mathrm{M}_a, \mathrm{M}_b, \alpha)$$

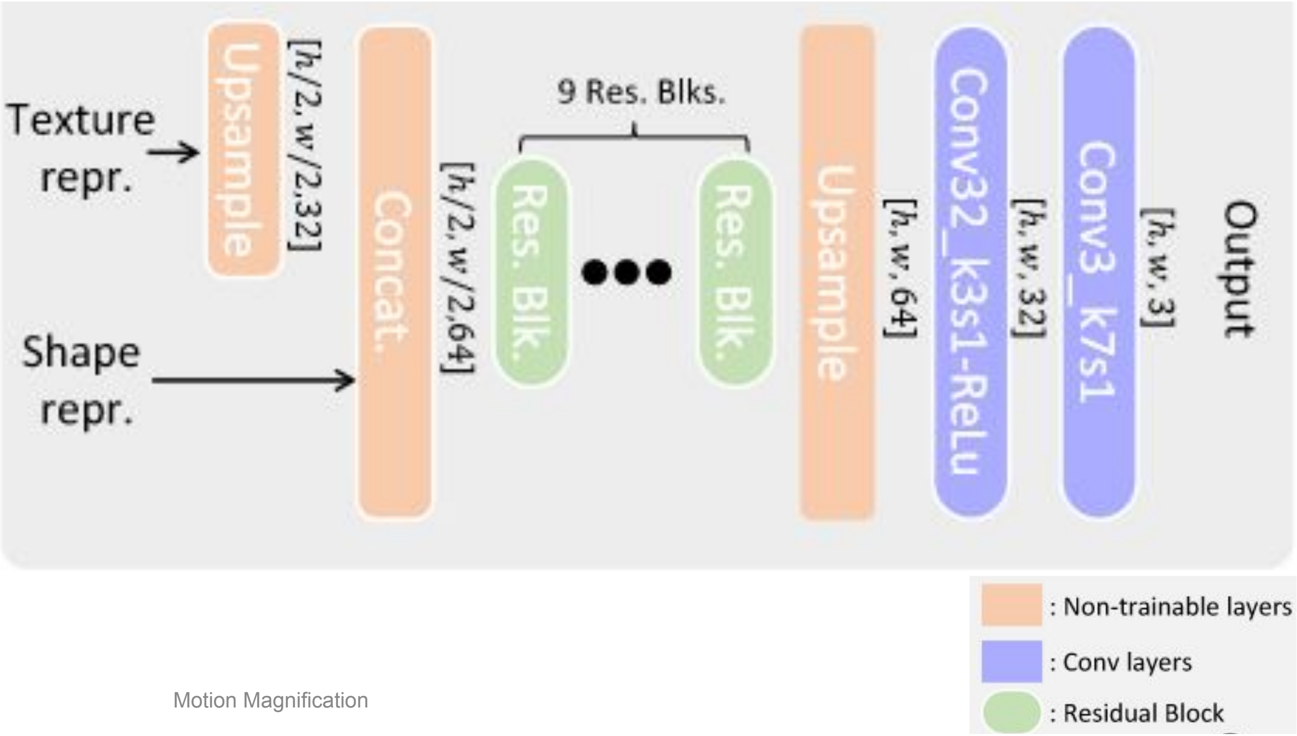


Manipulator - Temporal Filtering

$$G_{m,temporal}(\mathbf{M}(t), \alpha) = \mathbf{M}(t) + \alpha \mathcal{T}(\mathbf{M}(t)).$$

Where
M(t): Shape representation of the frame at time t
α: Magnification factor
τ(.): pixel-wise Temporal Filter

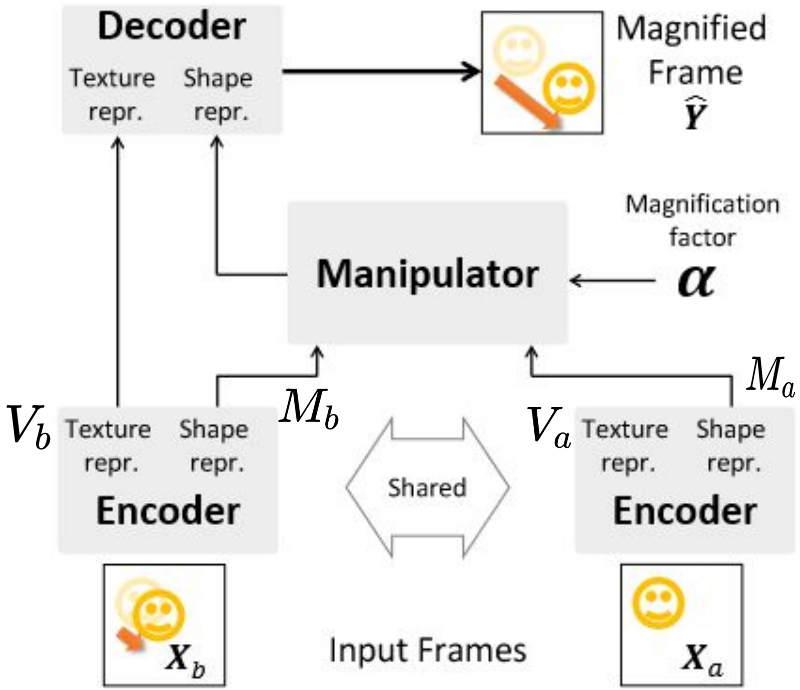
Decoder



Texture repr.
Upsample
[h/2, w/2, 32]
Shape repr.
Concat.
[h/2, w/2, 64]
9 Res. Blks.
Res. Blk.
Res. Blk.
Upsample
[h, w, 64]
Conv32_k3s1-ReLu
[h, w, 32]
Conv3_k7s1
[h, w, 3]
Output
: Non-trainable layers
: Conv layers
: Residual Block

Learning-Based Motion Magnification

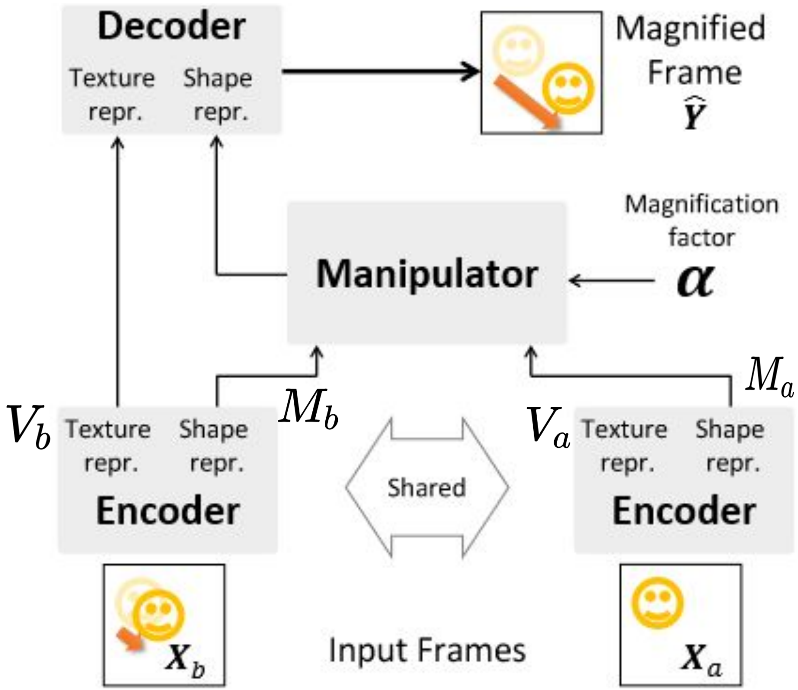
3.2. Deep Convolutional Neural Network - Minimization Problem



$$Loss = \left|\hat{Y} - Y\right| + \lambda(Regularisation)$$

Learning-Based Motion Magnification

3.2. Deep Convolutional Neural Network - Minimization Problem



$$Loss = \left|\hat{Y} - Y\right| + \lambda(Regularisation)$$

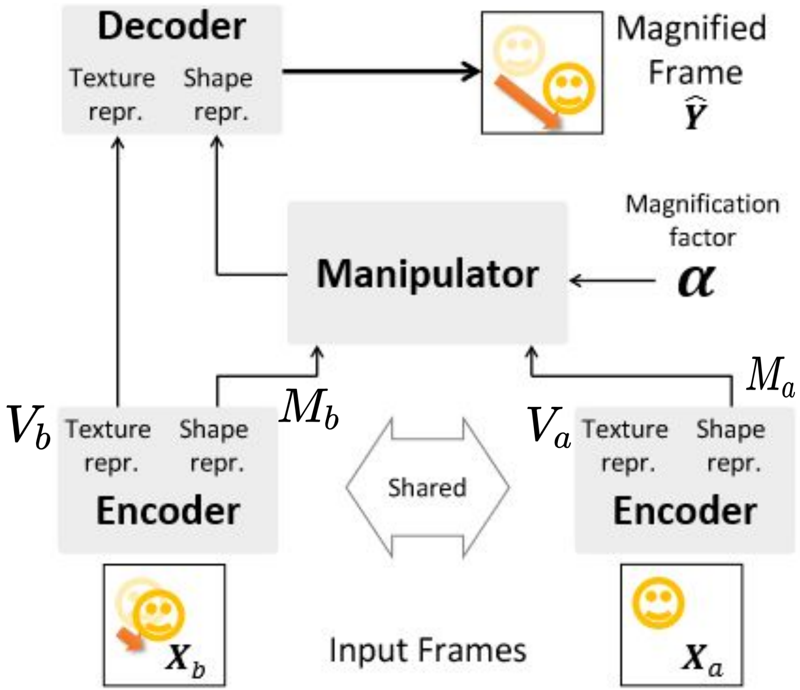
$$Y \;\; X_a \;\; X_b \longrightarrow X'_a \;\; X'_b \;\; Y'$$

color perturbation

- Shape representation should be invariant of color perturbation
- The texture representation of B should be close of Y
- The image A and B should have close texture representation

Learning-Based Motion Magnification

3.2. Deep Convolutional Neural Network - Minimization Problem



$$Loss = \left|\hat{Y} - Y\right| + \lambda(Regularisation)$$

$$Y \;\; X_a \;\; X_b \longrightarrow X'_a \;\; X'_b \;\; Y'$$

color perturbation

- Shape representation should be invariant of color perturbation
- The texture representation of B should be close of Y
- The image A and B should have close texture representation

$$Loss = \left|\hat{Y} - Y\right| + \lambda\left(|V_a - V_b| + \left|V'_b - V'_Y\right| + \left|M_b - M'_b\right|\right)$$

Results and evaluations

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Advantage of a **synthetic dataset** : large quantity

Foreground objects and background images: real image datasets for their realistic texture. 7,000 and 200,000 images respectively.

Each training sample: 7 to 15 foreground objects, randomly scaled from its original size. (the scaling factor limited at 2 to avoid blurry texture)

Amount and direction of motions of background and each object are randomized → learning local motions.

Low contrast texture, global motion, and static scenes: low performance => necessity of adding two types of example:

- the background is blurred
- only a moving background in the scene to mimic a large object

To learn changes due only to noise, addition of subsets of:

- Completely static scenes
- foreground moving but not background

.

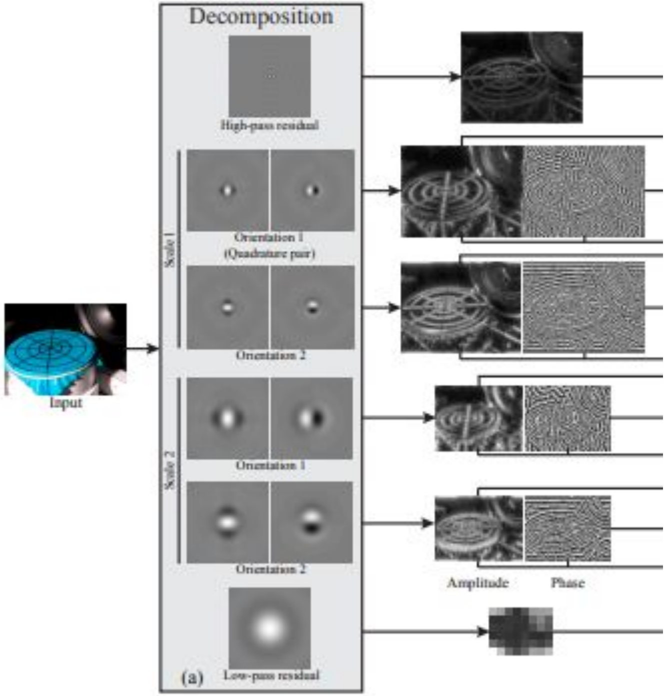
Subpixel motion generation: managing subpixel motion, it depends on demosaicking algorithm, so there is an approach with:

> **Reconstructing images in the continuous domain** using bicubic interpolation, before applying translation or resizing

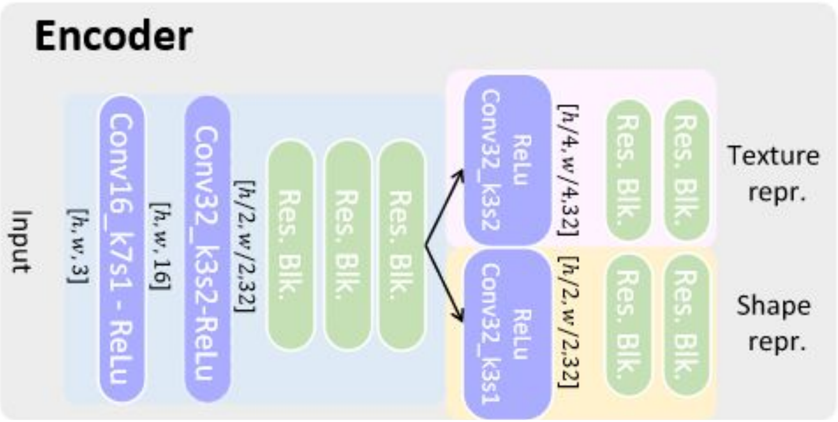
> **Generating images at a higher resolution** (motions appear larger) and then downsampling (to the desired size), with a Gaussian filter to reduce aliasing

> **Adding uniform quantization noise** before final quantization to ensure minor intensity changes are preserved: each pixel has a chance of rounding up proportional to its rounding residual

	Phase-based	Learning-based
First step	Decomposing with a Complex Steerable Pyramid	Encoder Ge
Second step	Filtering and Phase Removal Applying Motion Magnification	Manipulator Gm
Third step	Reincorporating High and Low Pass Components	Decoder Gd
Target	Phase of the steerable pyramid	Shape representation Ma
	Amplitude of the steerable pyramid	Texture representation Mb



Decomposition
High-pass residual
Scale 1
Orientation 1
(Quadrature pair)
Orientation 2
Scale 2
Orientation 1
Orientation 2
(a)
Low-pass residual
Input
Amplitude
Phase

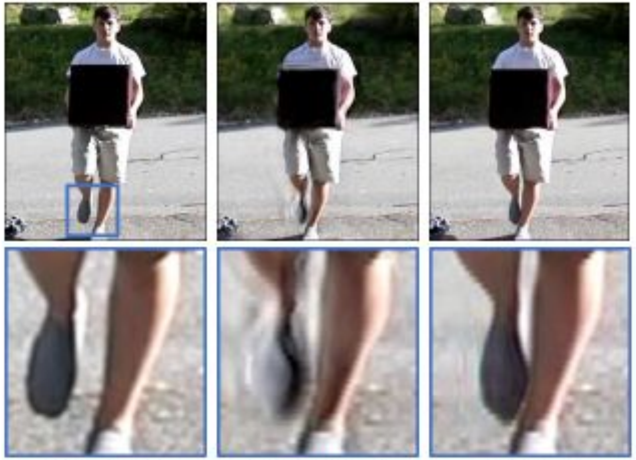


Encoder
Input
[h, w, 3]
Conv16_k7s1 - ReLu
[h, w, 16]
Conv32_k3s2-ReLu
[h/2, w/2, 32]
Res. Blk.
Res. Blk.
Res. Blk.
ReLu
Conv32_k3s2
[h/4, w/4, 32]
Res. Blk.
Res. Blk.
Texture repr.
ReLu
Conv32_k3s1
[h/2, w/2, 32]
Res. Blk.
Res. Blk.
Shape repr.

Results and Evaluations

4.2. Comparison of Two Approaches

(a) Phase (b) Ours



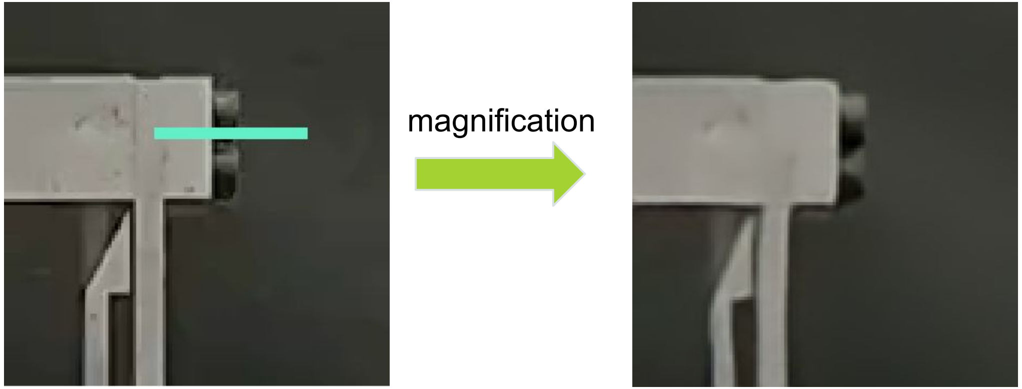
(c) Input (d) Phase (e) Ours

⇒ More ringing artifacts and blurring in the phase-based method than the learning based one.
⇒ Learning-based representation is trained **end-to-end** using example motion data. In contrast, the phase-based method relies on manually designed **multi-scale representations**, which struggle to accurately handle strong edges.

Results and Evaluations

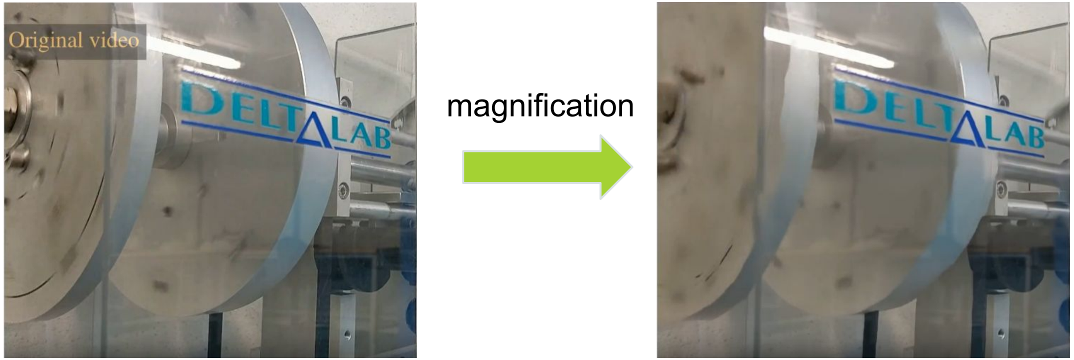
4.3. Comparison with state of the art

- Could not find benchmarks for motion magnification.
- However the Swin Transformer motion magnification model seems to be the most used.
- Different architecture, also trained on synthetiques training data



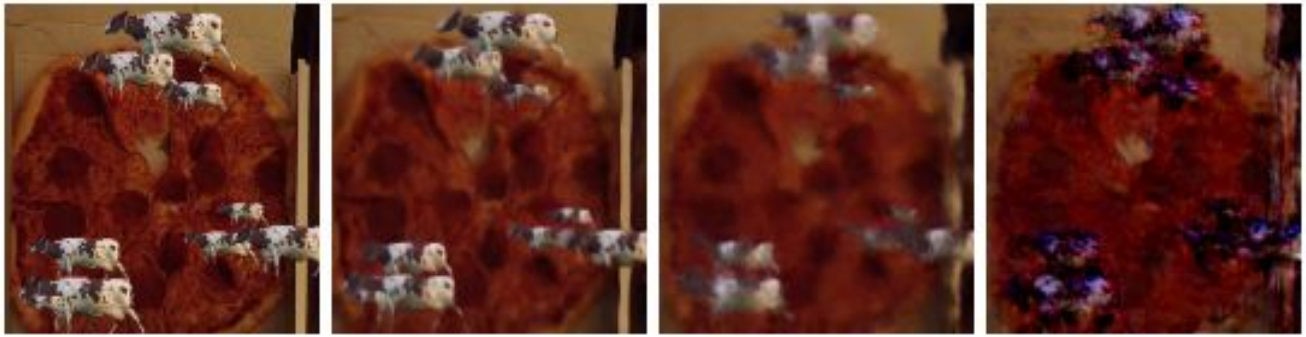
- Edge conservation
- Less artifacts

- Could not find benchmarks for motion magnification.
- However the Swin Transformer motion magnification model seems to be the most used.
- Different architecture, also trained on synthetiques training data s



- Edge conservation
- Less artifacts

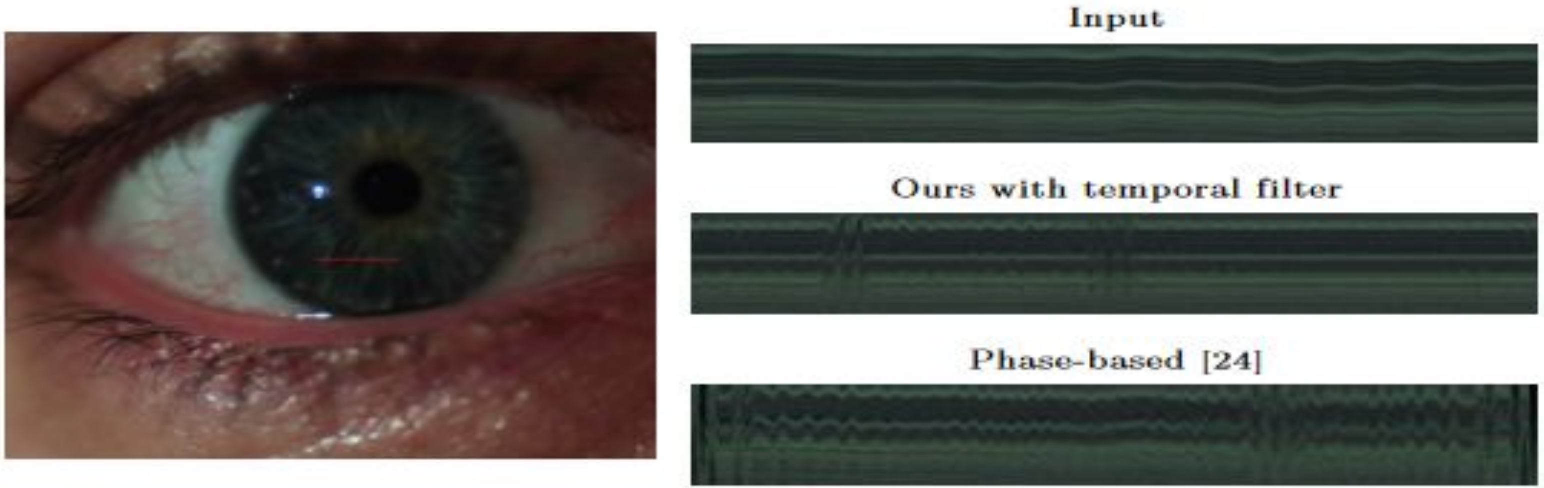
Performance degradation with high α



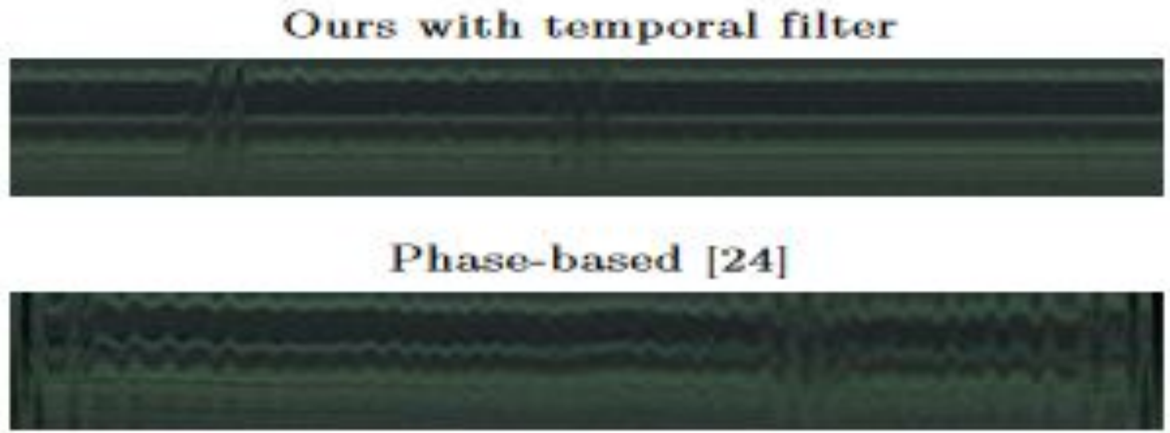
Original Frame 20× 50× 300×

Issue: Blurring and color artifacts with high magnification factors.

Blindness to very small motions



Blindness to very small motions



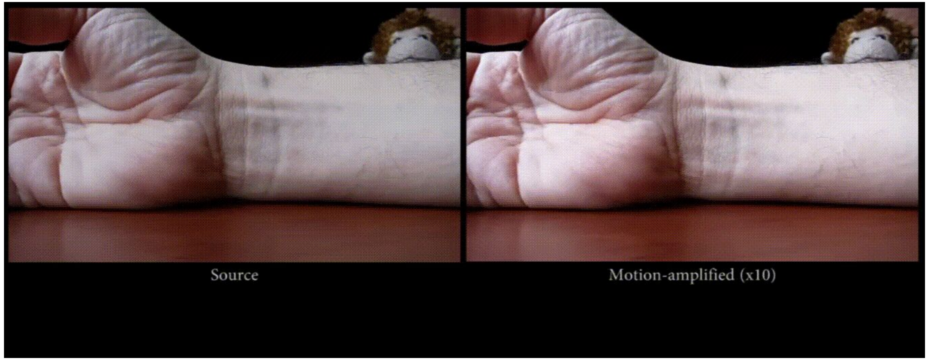
Issue: Occasionally magnifying the motion: *"Patchy Magnification"*
→ **Possible solution:** Improving compatibility with temporal filters.

Conclusion

Key takeaways from this lecture:

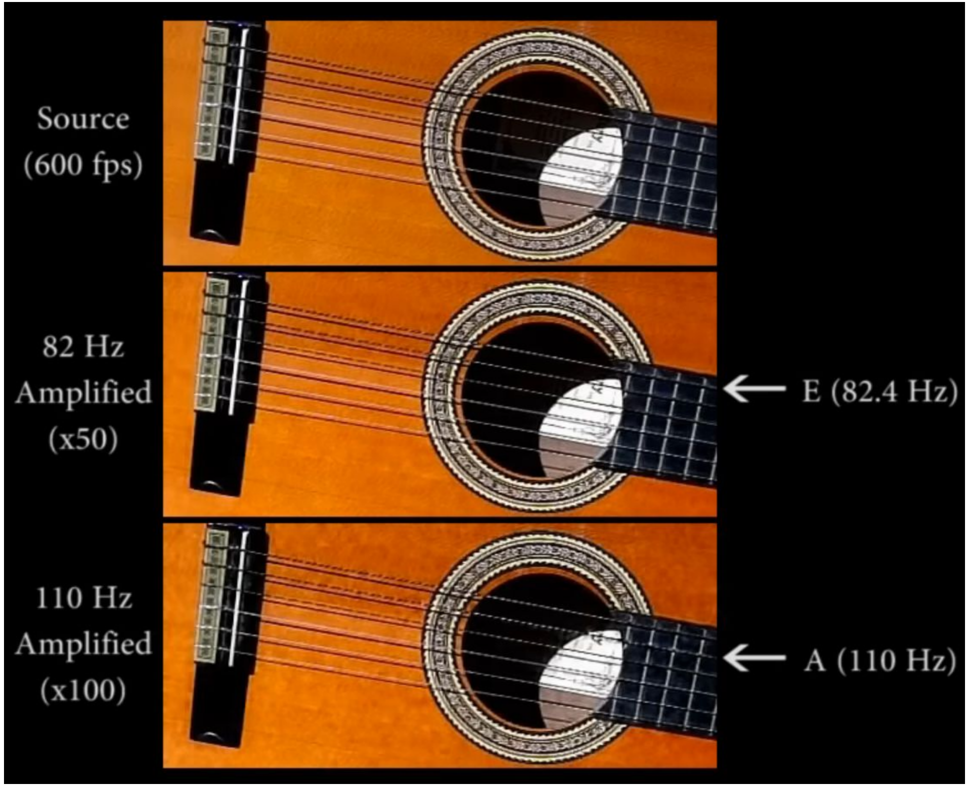
- **Phase-based approach:** Access to motion through the local phase of each sub band of fourier decomposition
- **Learning-based approach:** CNNs directly learn motion magnification filters from data.
- Both options present advantages and disadvantages.

- **Applications:**
- Medical applications: heart rate pulse measurement



Conclusion

- Predicting acoustic wave from the motion



Thank you for your attention

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Questions

This paper dates from 2018, what are the new approaches ?

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

The synthetic dataset that they designed contains a ground truth for each sample or is this an unsupervised problem? If these ground truths exist, what are they exactly?

What are some potential real-world applications of this motion magnification method?