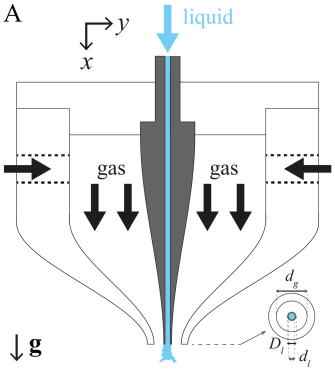
Generative AI for missing view generation in atomization experiments

Corentin Masson[1,2] Nathanaël Machicoane[1] Massih-Reza Amini[2]

[1]Univ. Grenoble Alpes, LEGI, CNRS
[2]Univ. Grenoble Alpes, LIG, CNRS

GdR TransInter – June 10, 2025

Atomization



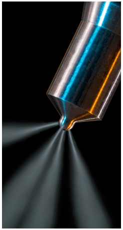
LEGI's coaxial two-fluid atomizer

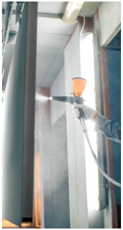
Gas-assisted atomization: liquid breakup by high-speed gas

Formation of a spray

Many applications:

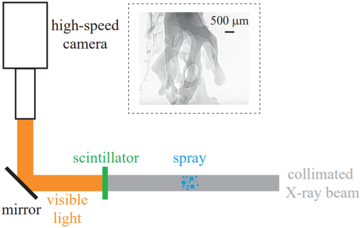
- Fuel injection
- Coating
- Spray drying
- Drug manufacturing





Project description

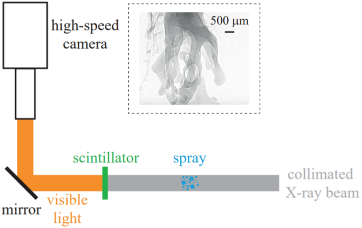
- Build the 3D model of the phenomenon
- Input: X-ray and visible light camera angles (projections)
- Rotation of the experiment prevented by Coriolis and centrifugal forces



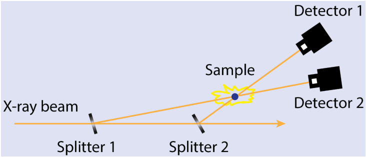
Machicoane et al. Synchrotron radiography characterization of the liquid core dynamics in a canonical two-fluid coaxial atomizer. (2019)

Project description

- Build the 3D model of the phenomenon
- Input: X-ray and visible light camera angles (projections)
- Rotation of the experiment prevented by Coriolis and centrifugal forces
- No beam splitter (too low quality)



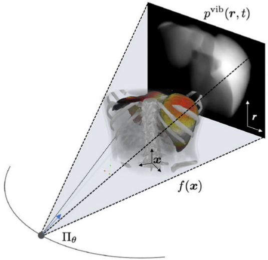
Machicoane et al. Synchrotron radiography characterization of the liquid core dynamics in a canonical two-fluid coaxial atomizer. (2019)



Zhang et al. 4D-ONIX: A deep learning approach for reconstructing 3D movies from sparse X-ray projections. (2024)

Project description

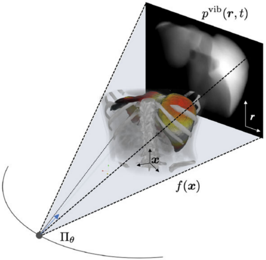
- Small number of projections → not only geometrical principles



Computed tomography usually require many projections.
Jailin et al. Measurement of 1–10 Hz 3D vibration modes with a CT-scanner. (2020)

Project description

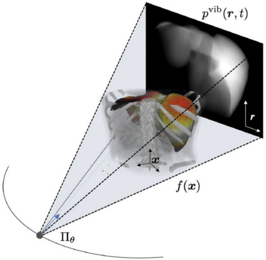
- Small number of projections $\rightarrow$ not only geometrical principles
- Help of generative AI



Computed tomography usually require many projections.
Jailin et al. Measurement of 1–10 Hz 3D vibration modes with a CT-scanner. (2020)

Project description

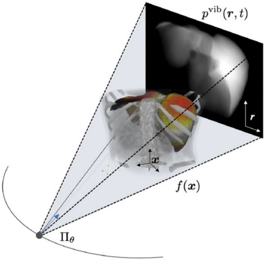
- Small number of projections → not only geometrical principles
- Help of generative AI
- Training data:
  - Real X-ray and visible light imaging (ESRF)
  - Simulation data



Computed tomography usually require many projections.
Jailin et al. Measurement of 1–10 Hz 3D vibration modes with a CT-scanner. (2020)

Project description

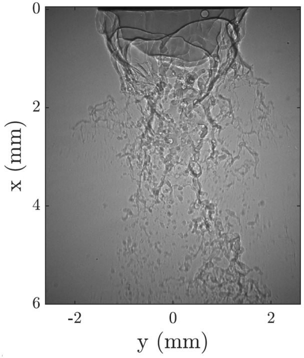
- Small number of projections $\rightarrow$ not only geometrical principles
- Help of generative AI
- Training data:
  - Real X-ray and visible light imaging (ESRF)
  - Simulation data
- Objectives:
  - Generate 1 novel view
  - Generate the full 3D structure
  - Achieve previous goals using as few input projections as possible

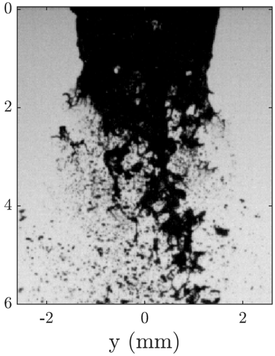


Computed tomography usually require many projections.
Jailin et al. Measurement of 1–10 Hz 3D vibration modes with a CT-scanner. (2020)

X-ray *vs* visible light imaging

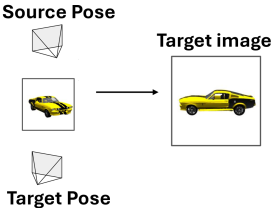
- Liquid $\leftrightarrow$ 1, gas $\leftrightarrow$ 0
- Spray $\mathcal{S} : \mathbb{R}^3 \to \{0, 1\}$
- View: projection on a $H \times W$ frame
- Visible light: $V_{visible} \in \{0, 1\}^{H \times W}$
- X-ray: $V_{Xray} \in [0, 1]^{H \times W}$





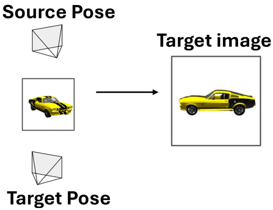
Missing view generation

- Active field of research in AI
- Recent developments
- From several input views $\{V_1, \ldots, V_n\}$, generate a novel view $\hat{V}_{n+1}$ (unseen pose)



Missing view generation

- Active field of research in AI
- Recent developments
- From several input views $\{V_1, \ldots, V_n\}$, generate a novel view $\hat{V}_{n+1}$ (unseen pose)
- Loss function:
  $MSE(V_{n+1}, \hat{V}_{n+1}) = \sum_{i,j}(V_{n+1}^{(i,j)} - \hat{V}_{n+1}^{(i,j)})^2$



True image

0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0

Predicted image

0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	1	1	0	0
0	0	1	0	1	0	0
0	1	0	0	0	1	0

MSE = 6

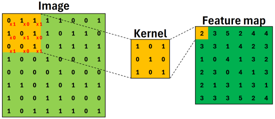
Convolutional Neural Networks (CNN)

- Original task: image classification

Convolutional Neural Networks (CNN)

- Original task: image classification
- Rely on convolutions with learnable filters

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

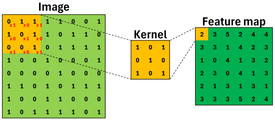


Convolutional Neural Networks (CNN)

- Original task: image classification
- Rely on convolutions with learnable filters

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

- Local receptive fields

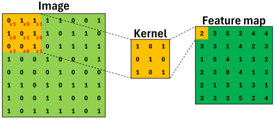


Convolutional Neural Networks (CNN)

- Original task: image classification
- Rely on convolutions with learnable filters

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

- Local receptive fields
- Shared weights

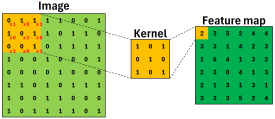


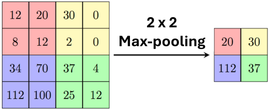
Convolutional Neural Networks (CNN)

- Original task: image classification
- Rely on convolutions with learnable filters

  $$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

- Local receptive fields
- Shared weights
- Pooling layers reduce spatial size





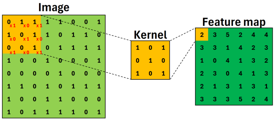
Convolutional Neural Networks (CNN)

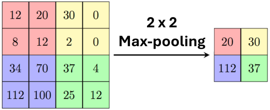
- Original task: image classification
- Rely on convolutions with learnable filters

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

- Local receptive fields
- Shared weights
- Pooling layers reduce spatial size
- Dense layer: $y_i = f(\sum_j W_{ij} x_j + b_i)$

W: weights - b: bias - f: activation function

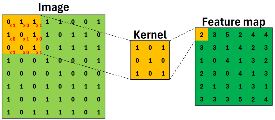




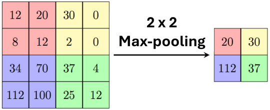
Convolutional Neural Networks (CNN)

- Original task: image classification
- Rely on convolutions with learnable filters

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n}$$

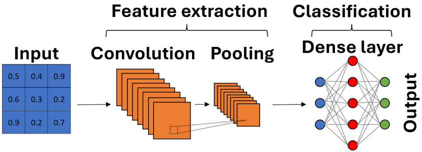


- Local receptive fields
- Shared weights
- Pooling layers reduce spatial size

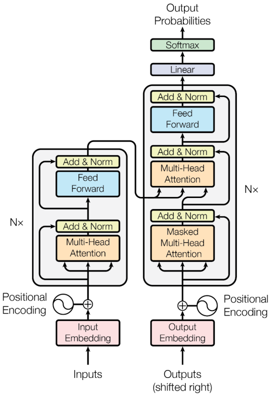


- Dense layer: $y_i = f(\sum_j W_{ij} x_j + b_i)$

W: weights - b: bias - f: activation function



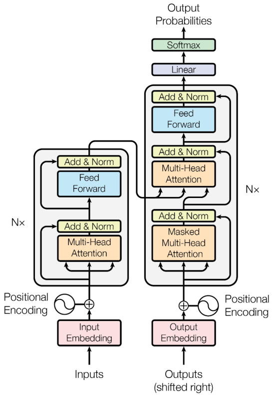
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)

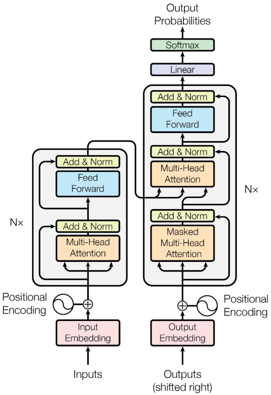
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder

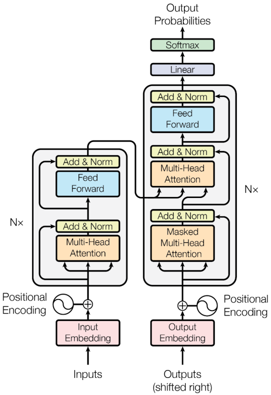
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions

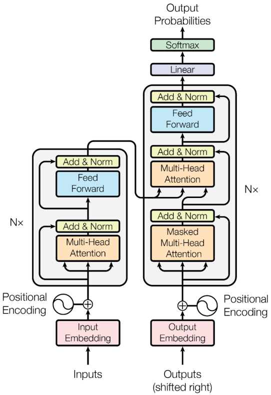
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions
- $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$

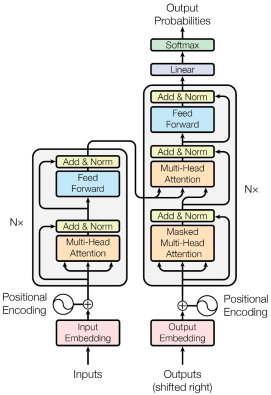
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions
- $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$
- $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$

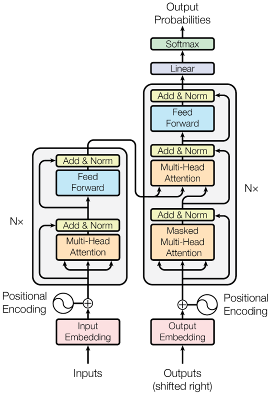
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions
- $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$
- $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$
- $QK^\top$: similarity between query and key

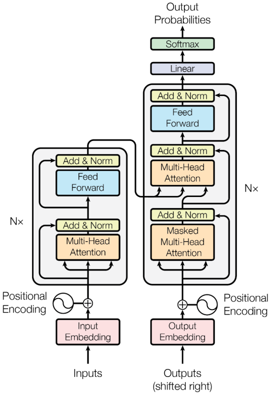
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions
- $\mathrm{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$
- $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$
- $QK^\top$: similarity between query and key
- Softmax converts similarities into a probability distribution

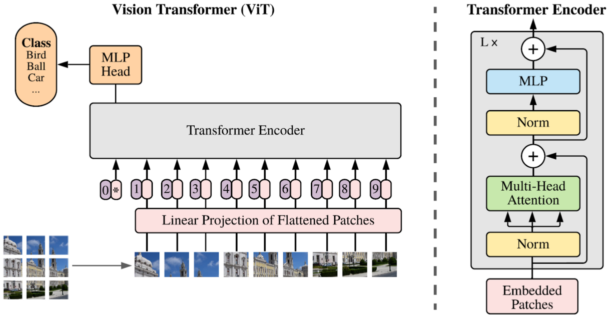
Classical transformers



Vaswani et al. Attention is All you Need. (2017)

- Original task: sequence-to-sequence (translation)
- Encoder + decoder
- Input: Embedded tokens + positions
- $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$
- $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$
- $QK^\top$: similarity between query and key
- Softmax converts similarities into a probability distribution
- GPT: Generative Pretrained Transformer

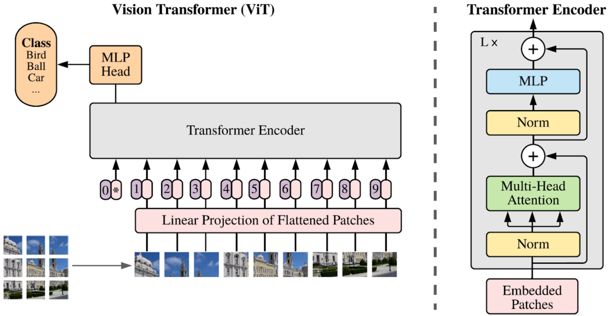
Vision transformers



Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021)

- Original task: image classification

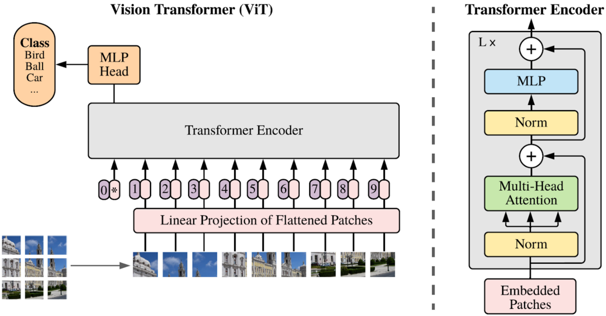
Vision transformers



Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021)

- Original task: image classification
- Transformer encoder with tokens being image patches

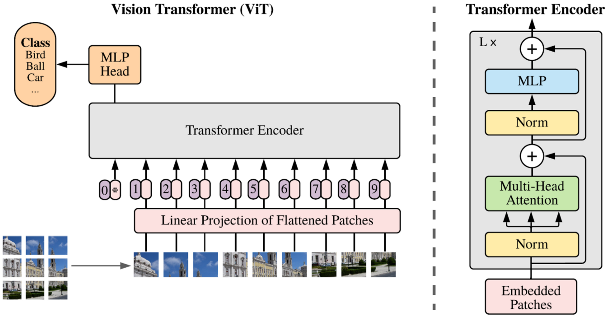
Vision transformers



Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021)

- Original task: image classification
- Transformer encoder with tokens being image patches
- Positional encoding

Vision transformers



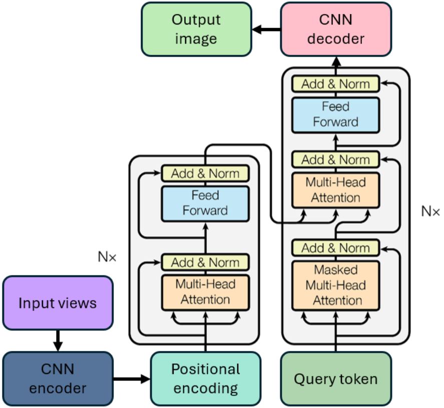
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021)

- Original task: image classification
- Transformer encoder with tokens being image patches
- Positional encoding
- For image reconstruction: decoder with learnable query token (geometrical information)

Proposed architecture

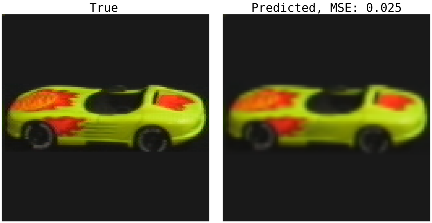
Vision transformer surrounded by CNN encoder and decoder

```
function CViT(x)                                  ▷ Input tensor of shape (V, H, W, C)
    for i = 1 to L_Ce do                          ▷ CNN encoder
        x ← Conv2D(x)
        x ← MaxPooling2D(x)
    end for
    pos ← Dense([1, ..., dim(x)])
    x ← x + pos
    for i = 1 to L_Te do                          ▷ Transformer encoder
        x ← SelfAttention(Q = x, K = x, V = x)
        x ← Dense(x)
    end for
    tokens ← x
    x ← Dense(Φ)                                  ▷ Geometrical information (camera angle Φ)
    for i = 1 to L_Td do                          ▷ Transformer decoder
        x ← SelfAttention(Q = x, K = x, V = x)
        x ← Attention(Q = x, K = tokens, V = tokens)
        x ← Dense(x)
    end for
    for i = 1 to L_Cd do                          ▷ CNN decoder
        x ← Conv2DTranspose(x)
    end for
    return sigmoid(x)
end function
```



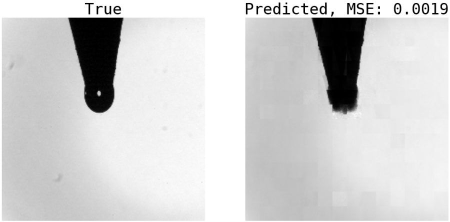
First results and perspectives

- Good results on benchmark datasets



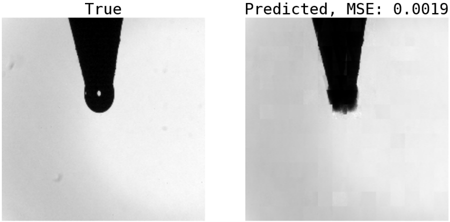
First results and perspectives

- Good results on benchmark datasets
- Poorer results on real images
- Only 2 input views (1 X-ray, 1 visible light)
- Patch artifacts, low quality



First results and perspectives

- Good results on benchmark datasets
- Poorer results on real images
- Only 2 input views (1 X-ray, 1 visible light)
- Patch artifacts, low quality



Future work:

- Increase the number of training images
- Add perceptual loss (loss over the features)
- Add physics-informed loss
- Train with simulation data
- Quantify the benefit of X-ray images
- Determine the number of views needed for good reconstruction

Thank You!

Questions?