UNIVERSITÀ DEGLI STUDI DI MILANO

Towards Explainable Semantic Segmentation For Autonomous Driving Systems By Multi-Scale Variational Attention

Mohanad Abukmeil[*], Angelo Genovese[*], Vincenzo Piuri[*], Francesco Rundo[‡], Fabio Scotti[*]

[*]Università degli Studi di Milano
Department of Computer Science
via Celoria 18, I-20133 Milano (MI), Italy
mohanad.abukmeil@unimi.it

[‡]STMicroelectronics
ADG
Central R&D, 95121 Catania (CT)
francesco.rundo@st.com

IEEE ICAS 2021 August 11-13, 2021

Outline

- Explainable autonomous driving systems (EADS)
- Semantic image segmentation
- Related works
- Proposed explainable variational attention
- Experimental results
- Recent works comparison
- Conclusions

EADS

- In ADS, vehicles can sense the surrounding environment to perform driving tasks
  - Control engine
  - Visualize objects
  - Anomaly detection, etc.

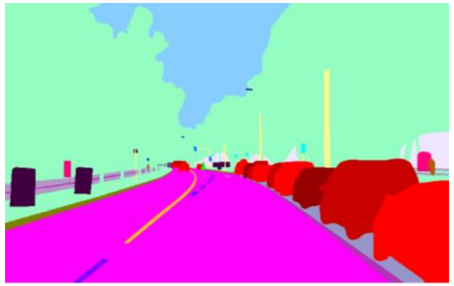
- Driving tasks can be learned by MLs
  - Automatically processing data
  - Recognizing objects
  - Instantaneous recommendations
- Explainable artificial intelligence (XAI) explains the behaviors and decisions of the MLs

EADS and semantic segmentation

- EADS combine XAI and ADS to enhance the vehicular automation (VA)
  - Interpreting sensory data
  - Mentoring vehicles behaviors
  - Semantically segmenting the ambient Objects
- In explainable semantic segmentation, each pixel holds a semantic meaning
  - Explains detected objects
  - Offers road conditions

Image

Mask



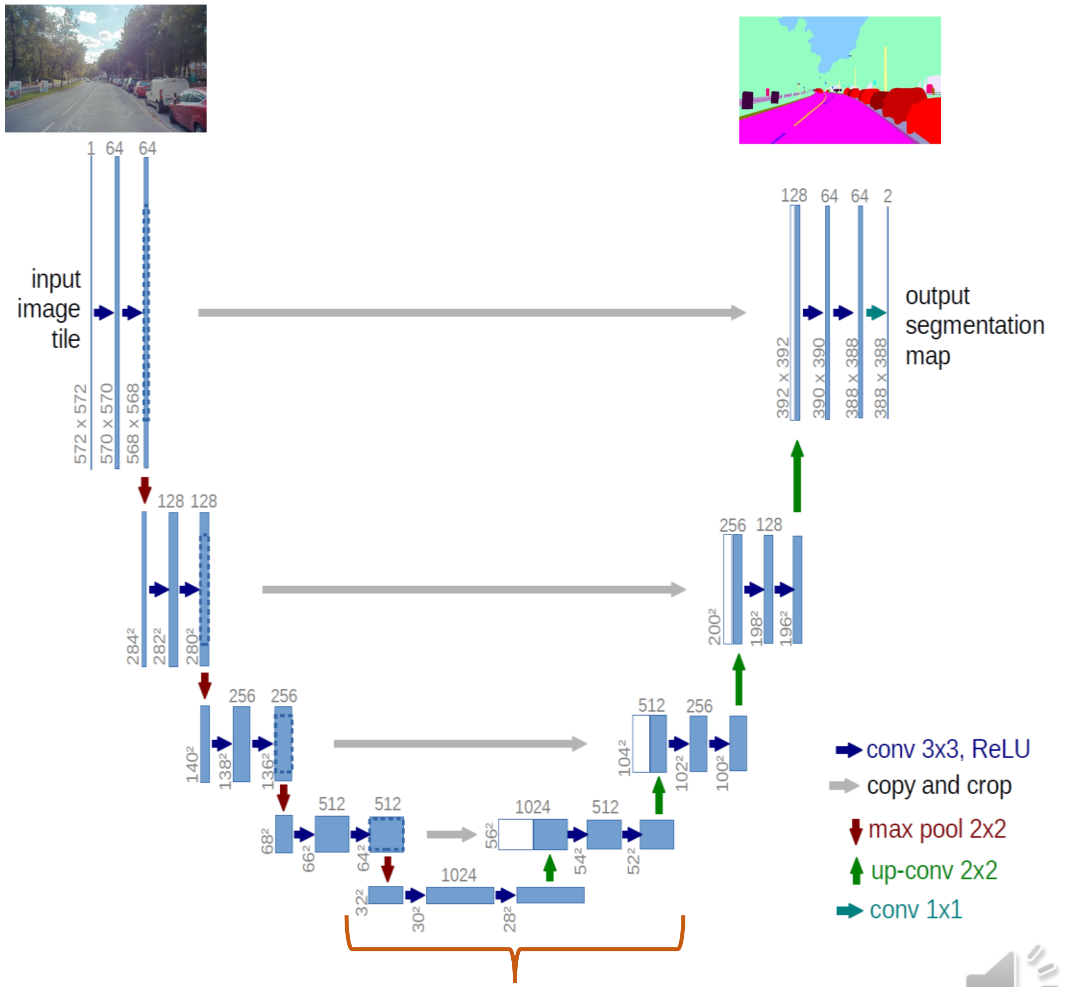
Related semantic image segmentation works

- The success of AEs led to the flourishing of segmentation models
- Two main stages (blocks)
  - Encoder maps images to latent space $f: \mathbb{R}^D \longrightarrow \mathbb{R}^d,\ d << D$
  - Decoder reconstructs segmented masks $g: \mathbb{R}^d \longrightarrow \mathbb{R}^D$
- All segmentation models optimize a similar objective
  - Irrespective of the error metric

$$\mathbf{L}_{\mathrm{rec}_{\{\hat{\theta}_e,\, \hat{\theta}_d\}}} = \min \|\underset{\text{GT mask}}{X} - \underbrace{(f \circ g)X}_{\text{Reconstructed mask}}\|^2_{\mathrm{Er}}$$

Related semantic segmentation works

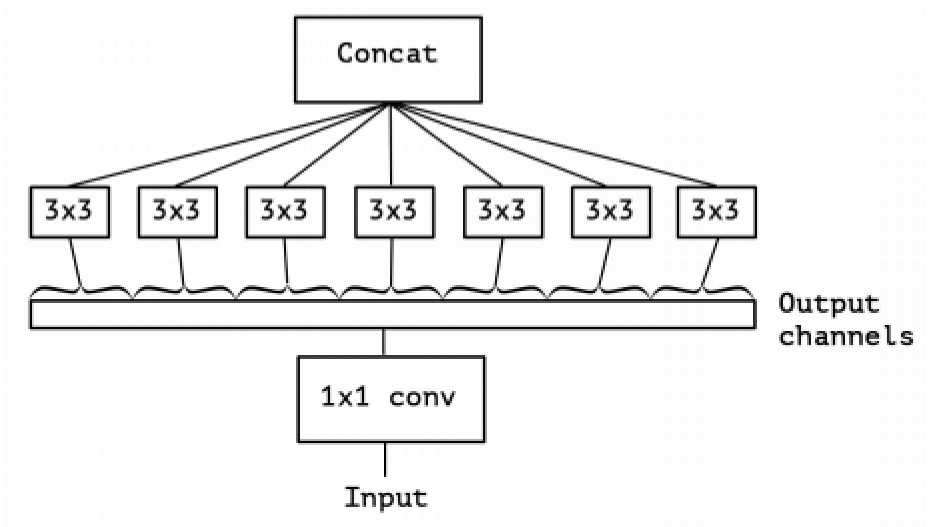
- **U-net** performs a typical AEs mapping for image segmentation
  - Blue boxes are feature channel
- Concept: copy and concatenate blocks
  - Training similarly to AEs
  - Convolutional block in the latent space



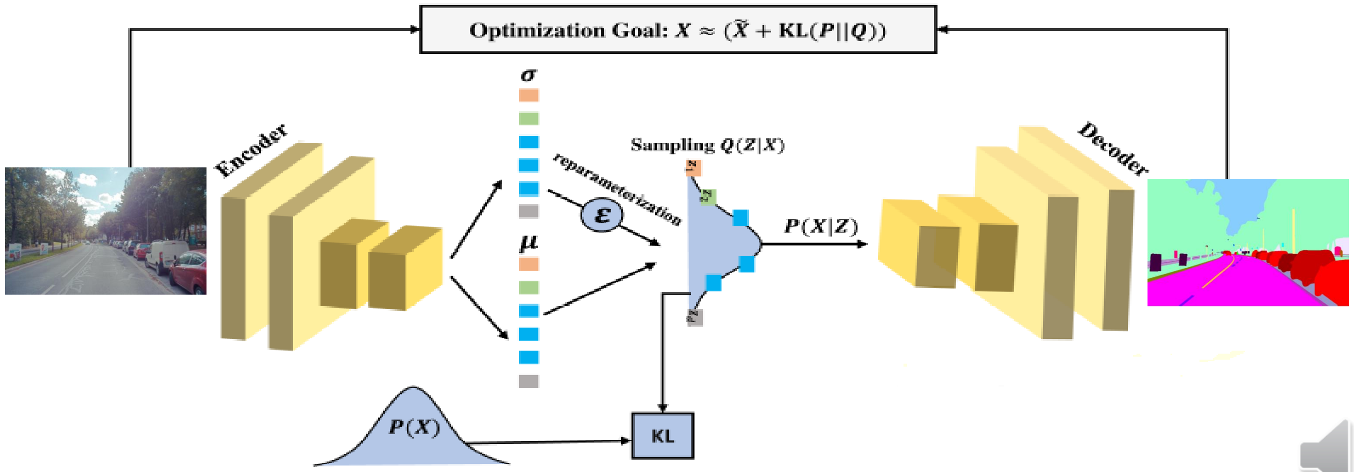
Latent Space

Related semantic segmentation works

- **Xception-based U-Net**
  - Uses separable depth convolution
  - Utilizes residual learning $X + f(X)$
  - Employs inception blocks for each layer



- **Deep VAE**
  - It is regularized by VI
  - Optimizes two losses
  - FCL layer in the latent space



Proposes methodology (1/2)

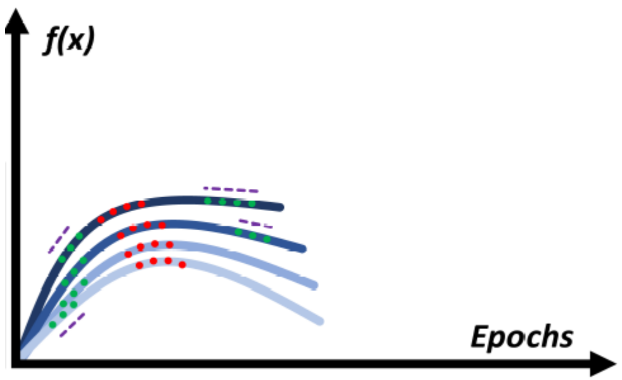
- The performance of recent DL models is limited
  - Traditional optimization objective
  - Lack of the explainability
  - Architecture complexity
- $\mathbf{Mgrad_2}$**VAE**: A novel variational segmentation model for EADs
  - Regularized by VI
  - Uses the second-order partial derivative to build an attention map, $\frac{\partial^2 Z}{\partial L_i^2}$
  - Offers online (learnable attention) for each encoding layer
  - It is optimized based on VAE loss + attention loss simultaneously

Proposes methodology (2/2)

- The second-order partial derivative capture variation of the gradient
  - The variation reflects the curvature of the activation functions
  - Each latent neuron gives specific activation respecting the encoder layers

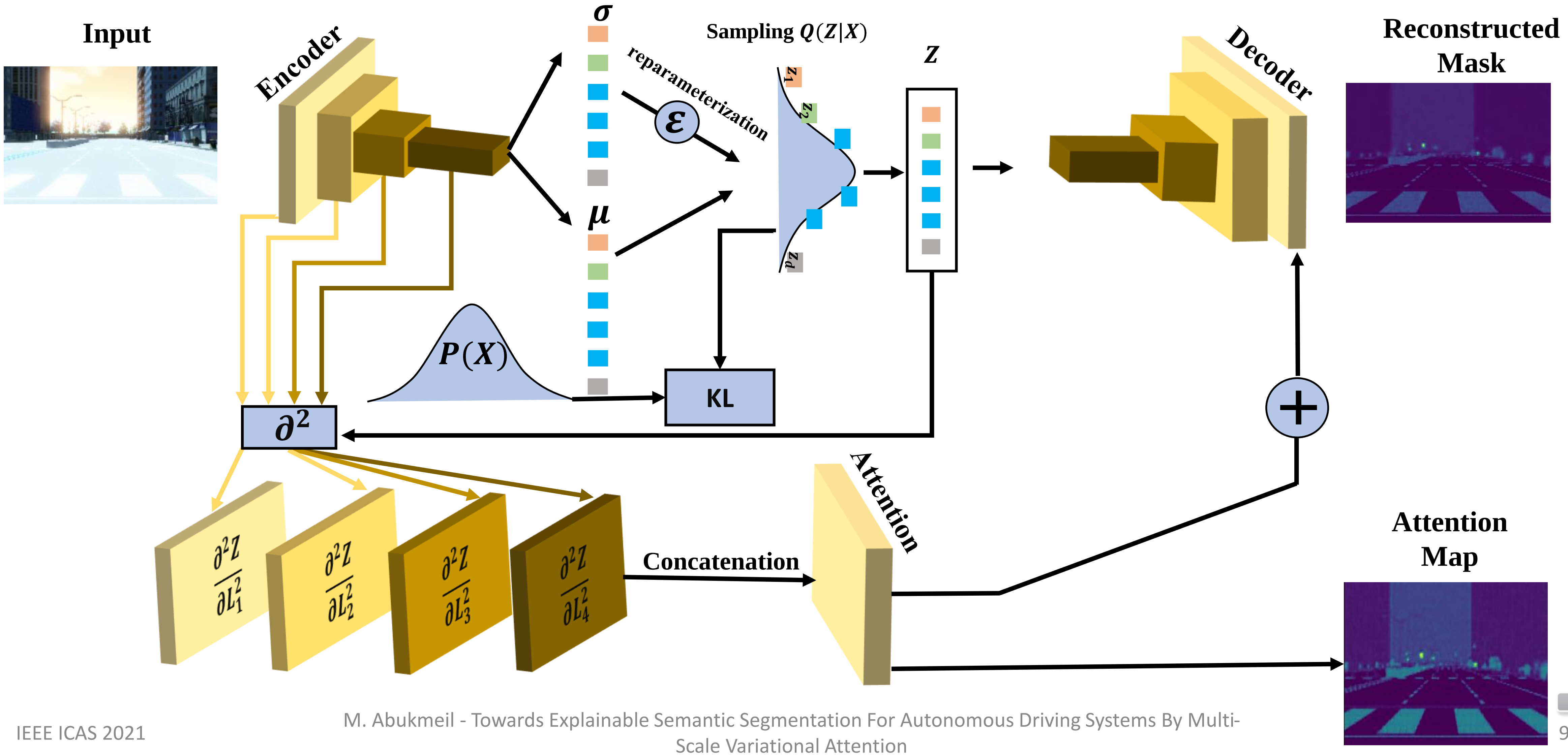
$$\text{Scale}_1 : \frac{\partial^2 Z}{\partial L_1^2},$$

$$\text{Scale}_l : \frac{\partial^2 Z}{\partial L_l^2},$$



- Aggregate all second-order partial derivative tensors
  - Summation $\left(\sum_{i=1}^{l} \frac{\partial^2 Z}{\partial L_i^2}\right)$
  - Average $\left(\frac{1}{l} \sum_{i=1}^{l} \frac{\partial^2 Z}{\partial L_i^2}\right)$
  - Convolutional layer (Concat+Conv)

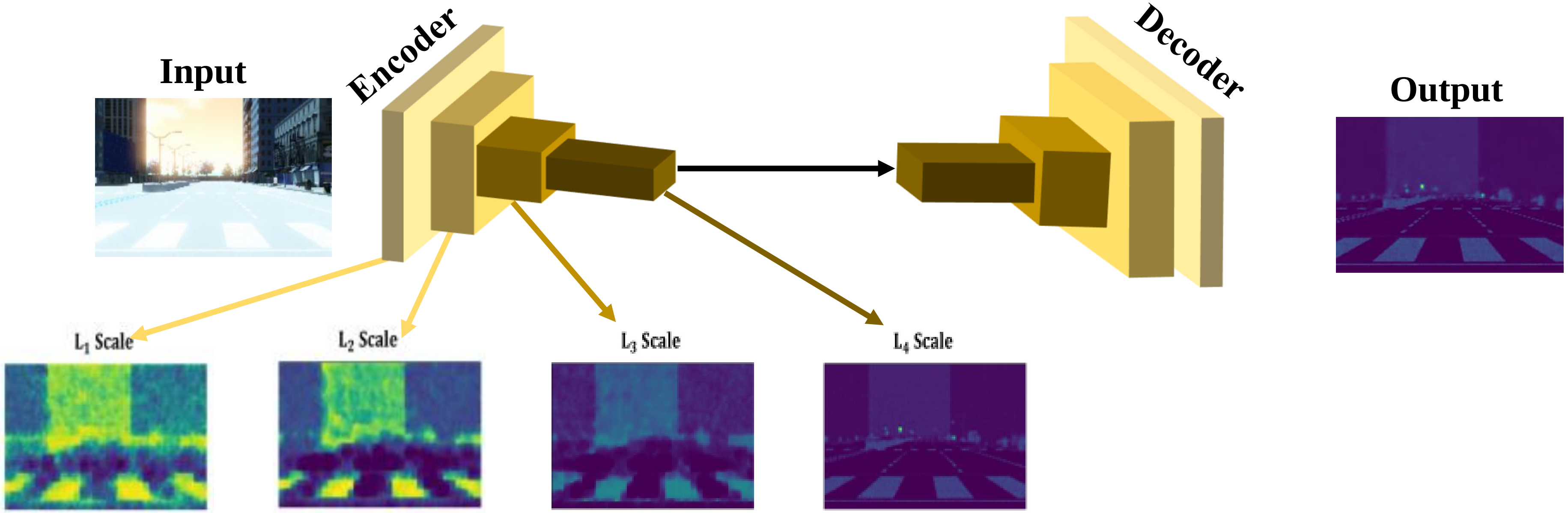
Mgrad_2VAE



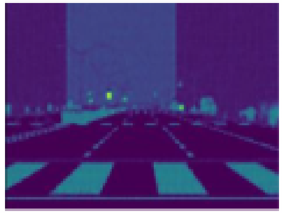
Mgrad_2VAE optimization and performance evaluation

- DVAE loss: $$\mathbf{L}_{\theta_{\text{VAE}}} = \min[\mathbf{L}_{\text{rec}} + \text{KL}(P \| Q)]$$
- Mgrad_2VAE: $$\mathbf{L}_{\text{Mgrad}_2\text{VAE}} = \min[\underbrace{\mathbf{L}_{\text{VAE}}}_{\text{DVAE loss}} + \underbrace{\|X - \theta_{\text{Mgrad}}(Z, L_{e_i})\|^2_{\text{Er}}}_{\text{Attention loss}}]$$
- SYNTHIA and A2D2 datasets have been considered
- Qualitative and Quantitative analysis used
  - SSIM metric used in qualitative analysis
  - AUC-ROC metric used in quantitative analysis

Mgrad$_2$VAE optimization and performance evaluation



- Final attention map



SSIM Index	SYNTHIA	A2D2
Reconstructed masks	97.57%	60.38%
Attention maps	96.47%	55.71%

Experimental results: Quantitative comparison with the literature

- Pixel-wise predictive performance comparison with recent models
  - Methods in the literature:
    - DVAE
    - Xception model built based on the U-net architecture (transfer Learning)
  - Proposed model: $\mathbf{Mgrad_2}$**VAE**

AUC-ROC	SYNTHIA	A2D2
Deep VAE	79.60%	94.05%
Xception	67.43%	95.19%
Our $\mathbf{Mgrad_2}$VAE reconstruction	**81.50%**	**95.44%**
Our $\mathbf{Mgrad_2}$VAE attention	**83.20%**	**95.36%**

Conclusions

- First ESS model for EADS
- Second-order derivative to capture the curvature of neuron activations
  - An attention map for each encoding layer (multiscale)
- Online attention loss to improve the segmentation accuracy
  - Based on the residual fusion between the attention and the reconstructed mask
- High performance
- **In future works, we plan:**
  - Investigate XAI potential in harsh environment
  - ESS under rough weather conditions

thank you

Mohanad Abukmeil

https://homes.di.unimi.it/abukmeil/

Mohanad.abukmeil@unimi.it