TOWARDS EXPLAINABLE SEMANTIC SEGMENTATION FOR AUTONOMOUS DRIVING SYSTEMS BY MULTI-SCALE VARIATIONAL ATTENTION

Mohanad Abukmeil‡, Angelo Genovese‡, Vincenzo Piuri‡, Francesco Rundo†, and Fabio Scotti‡
‡ Department of Computer Science, Università degli Studi di Milano, Italy {*firstname.lastname*}@unimi.it
† STMicroelectronics, ADG, Central R&D, 95121 Catania (CT), Italy *francesco.rundo*@st.com

ABSTRACT

Explainable autonomous driving systems (EADS) are emerging recently as a combination of explainable artificial intelligence (XAI) and vehicular automation (VA). EADS explains events, ambient environments, and engine operations of an autonomous driving vehicular, and it also delivers explainable results in an orderly manner. Explainable semantic segmentation (ESS) plays an essential role in building EADS, where it offers visual attention that helps the drivers to be aware of the ambient objects irrespective if they are roads, pedestrians, animals, or other objects. In this paper, we propose the first ESS model for EADS based on the variational autoencoder (VAE), and it uses the multiscale second-order derivatives between the latent space and the encoder layers to capture the curvatures of the neurons' responses. Our model is termed as Mgrad$_2$VAE and is bench-marked on the SYNTHIA and A2D2 datasets, where it outperforms the recent models in terms of image segmentation metrics.

Index Terms**—** Autonomous Driving System, VAE, XAI, ESS.

1. INTRODUCTION

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has lead to the development of AI-powered autonomous systems, which can sense, learn, decide and interact for many different applications including computer vision, natural language processing (NLP), robotics, autonomous driving, and others fields[1, 2]. Moreover, AI-powered autonomous systems are build based on deep learning (DL) models comprising convolutional neural networks (CNN), autoencoders (AEs), generative adversarial networks (GANs), and Bayesian models [3]. However, the effectiveness of many recent models and systems are limited due to the scarcity of explainability; such an explainability translates the actions and decisions of the learned models to users who operate and develop them. Explainable artificial intelligence (XAI) is a branch of AI aims to explain the behaviors of the ML models [4].

An autonomous driving system (ADS) is referred to any vehicle that can sense the surrounded environment without human control, or with a limited level of supervision. ADS is also able to control engines, visualize objects, detect abnormal actions, drive vehicles, and activate breaks [5]. AI and ML influence ADS by automatically processing data, offering instantaneous recommendations, and recognizing objects; such objects include pedestrians, trees, bicyclers, and other moving and static objects [6, 7]. Explainable autonomous driving systems (EADS) combine XAI and ADS to enhance the vehicular automation (VA), throughout interpreting sensory data, mentoring vehicles behaviors, and semantically segmenting the ambient objects [4]. In this regard, the explainable semantic segmentation (ESS) is a branch of the ML in which each pixel of the segmented object holds a semantic meaning, and can be integrated into the EADS to improve the explainability of the detected objects, and to offer roads conditions conclusion to the drivers [8].

XAI-powered models are associated with unsupervised learning (UL) to visualize the hidden structure of data [9, 1]. AEs are a class of UL methods that are able to generate and visualize data, reduce dimensionality, and perform other ML tasks such as object recognition [10]. AEs comprise classic, de-noising, contractive, sparse, variational-AE (VAE) [10, 11]. Moreover, the success of AEs architectures led to the flourishing of different supervised AEs for structured prediction, i.e., semantic segmentation, such as Seg-net, U-net, and others [12, 13, 14]. Among all AEs, VAE is regulated by the variational inference (VI) to optimize the posterior distribution of large datasets, which leads to a better generalization. The VAEs have been utilized in the ADS in the absence of XAI, where it has been used in the steering control [15], pedestrian prediction in [16], trajectory simulation in [17], and anomaly detection for ADS [18].

The first work towards explaining the VAE behavior is proposed in [19], where it generates visual attention to show how the encoder side behaves. Moreover, the proposed attention map is built by duplicating the last layer of the encoder, thereafter it scales each feature point in the filter channels by a global average pooling of the gradient of the latent space concerning that layer. Factually, the drawback of such attention lies in the unfair scaling, i.e., both related and unrelated feature points are scaled with the same factor. On the other hand, the first work that has been attempted to build the attention of the CNN for ADS is described in [6], where the attention is built by averaging the activations of 100 images; such attention hides the effects of the high and low activations, i.e., approximated attention, and is not stable for time-series segmentation.

To fill the gap of explaining VAEs in the EADS applications, we propose Mgrad$_2$VAE[1], a novel ESS model for EADS applications. Moreover, the Mgrad$_2$VAE utilizes the multiscale second-order derivative between the latent space and each encoder layer, which captures the curvatures of neurons' activations to build the multiscale explainable attention without unfair scaling or averaging the final attention. Therefore, our contribution is twofold: *(i)* introducing a novel ESS model for EADS applications, by using the unsupervised VI and a supervised convolutional AE, and *(ii)* proposing a novel multi-scale gradient attention mapping scheme for ESS to improve EADS applications using the second-order derivative operator. The rest of this paper is organized as follows. Section 2 highlights the VAE and the proposed explanation methodology. Section 3 describes the architecture of Mgrad$_2$VAE. The experimental results are given in Section 4. The conclusion and future works are reported in Section 5.

This work was supported in part by the EC within the H2020 Program under projects MOSAICrOWN and MARSAL and by the Italian Ministry of Research within the PRIN program under project HOPE. We thank the NVIDIA Corporation for the GPU donated.

[1] The source code is available at:
`http://iebil.di.unimi.it/mgradvae/index.htm`

2. VAE AND THE EXPLAINABILITY METHODOLOGY

2.1. VAE

VAEs consist of many different encoding and decoding stages, where each stage represents a different scale of dimensionality that is contracted or expanded by using learning parameters θ (where $\theta = \{W, B\}$, W and B are weights and biases, respectively) [17]. The learning parameters are used to perform many different mapping including convolution, dense multiplication, deconvolution, regularization, etc, by utilizing several sets of representations to capture neurons' activations [20]. Also, for each setting among parameters θ, i.e., after each learning epoch, the gradient of the output is estimated with respect to the input by employing the first-order (1st) partial derivative to optimally reconstruct or generate data.

VAE encompasses two main modules [11]: *(i)* the inference (encoder) module that is used to map an image (or data) $X = \{x_i \,|\, x_i \in \mathbb{R}^D,\, i = 1, \dots, N\}$, D is the original dimensionality ($D = m \times n \times c$ which indicates rows, columns, and channel depth, respectively), to a latent space $Z = f(X) = \{z_i = f(x_i) \in \mathbb{R}^d, \,|\, i = 1, \dots, M\}$. Moreover, the encoder module reduces dimensionality of the data, i.e., $0 < d < D$, and it is used to infer the model likelihood $P(X|\theta)$ [10]. *(ii)* The generation (decoding) module that is utilized to reconstruct the original data $\tilde{X}$ from the latent space Z. For a given data $X \in \mathbb{R}^D$, the encoding module creates a mapping $f : \mathbb{R}^D \to \mathbb{R}^d$, while the decoding module creates an inverse mapping $g : \mathbb{R}^d \to \mathbb{R}^D$, which generates an approximation of the data: $\tilde{X} = g(Z; \hat{\theta}_d)$ [21]. Similarly to AEs, VAE is regulated to find the optimal set of parameters $(\hat{\theta}_e, \hat{\theta}_d)$ that achieve a better generalization [14], and to attain the minimum reconstruction loss $\mathbf{L}_{\text{rec}}$:

$$\mathbf{L}_{\text{rec}_{\{\hat{\theta}_e,\, \hat{\theta}_d\}}} = \min \|X - (f \circ g)X\|^2_{\text{Er}} \tag{1}$$

where E_r represents the reconstruction error metric which can be computed by reconstruction cross-entropy, β- divergence, mean square error (MSE), Frobenius norm, or β- divergence [9].

VI is utilized to regulate the VAE, where two different losses are optimized simultaneously for a better generalization [11]. The VI is a Bayesian method that approximates an intractable posterior over a large dataset, throughout approximating the probability densities by optimization. The VAE's encoder approximates the posterior distribution $Q(Z|X)$, which identifies the distributional shape of the latent space Z according to the original data X. Moreover, the VAE is characterized by $Q(Z|X)$ optimization; such an optimization affects the distribution of latent space Z to follow a Gaussian distribution with a definite mean μ (which reflects the Gaussian's center), and standard deviation σ (which reflects the Gaussian's shape).

Practically, the prior distribution of the latent space $P(Z)$ is considered (simply by duplicating the unit Gaussian distribution of th original data manifold $P(X)$); subsequently, the prior $P(Z)$ and the approximated distribution $Q(Z|X)$ are matched by utilizing the KL divergence [22]. The KL divergence is always positive and tends to zero if and only if P and Q are almost equal in the distribution, and it is mathematically defined as $\text{KL}(P\|Q) = \Sigma_x P(x) \log \frac{P(x)}{Q(x)}$. The variational process is known as the reparameterization trick, and it can be obtained by perturbing σ with a small noise ϵ, thereafter directing the optimizer to enforce the AE to reconstruct the data concerning the distribution of X. Moreover, the reparameterization trick augments the generalization, where it produces different distributions to be compared with $P(Z)$ as in duplicating data [11].

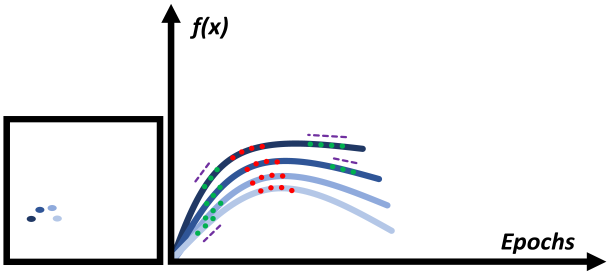

Fig. 1: The neurons activations and gradient over epochs.

Eventually, the VAE optimizes the reconstruction loss $\mathbf{L}_{\text{rec}}$ through minimization in accordance to Eqn. (1), and it is also optimized to minimize the distributional loss of the latent space between $Q(Z|X)$ and $P(Z)$ using $\text{KL}(P\|Q)$, that reflects which extent the reparameterized latent distribution follows a unit Gaussian:

$$\mathbf{L}_{\theta_{\text{VAE}}} = \min[\mathbf{L}_{\text{rec}} + \text{KL}(P\|Q)] \tag{2}$$

where $\theta_{\text{VAE}} = \{\hat{\theta}_e,\, \hat{\theta}_d,\, \hat{\mu}_X,\, \hat{\sigma}_X,\, \hat{\mu_Z},\, \hat{\sigma_Z}\}$.

2.2. The explainability methodology

Deep semantic segmentation models comprise many different encoding and decoding blocks to map data from the domain of the original image to the corresponding masks [12, 13]. Moreover, neurons with different parameters (θ_e, θ_d) are employed to optimally fit models, where at each learning epoch the gradient that measures the instantaneous rate of change among the model parameters is measured, by utilizing the first-order partial derivative ∂ between each pixel in the segmented mask with respect to the input image [23].

Considering a VAE with a single encoding layer L_{e1} and a latent layer Z, the first gradient between Z and L_{e1} is estimated according the partial derivative of each neuron activation z_i as $\frac{\partial z_i}{\partial L_{e1}}$. Moreover, if an additional layer L_{e2} lies between L_{e1} and Z, then the chain rule is used as $\frac{\partial z_i}{\partial L_{e1}} = \frac{\partial z_i}{\partial L_{e2}} \frac{\partial L_{e2}}{\partial L_{e1}}$ [24]. The result of all derivations gives the required rate of changes to update θ. Given a period of time, the neuron activations are changing; capturing such variations draws an attention map that gives an insight into how the neurons respond among different inputs, and it is obtained by considering the derivative of the gradient, i.e., 2nd partial derivative $\frac{\partial^2 z_i}{\partial L_{e1}^2}$ [25].

Visually, four pixels of an image with their associated neurons activations are illustrated in Fig. 1, where the activations are given according to the non-linear ReLU functions [26] (the method is valid for other types of activations). The 1st gradient is the slope (magenta dashed lines) at any point in the curves (blue curves), where the derivative of the gradient interprets how the curves are varied during a time (the red and green points). As it is observed from Fig. 1, the gradient of activations can be stationary during a period of the learning time, i.e., the 2nd derivative around the green points is ≈ 0, however, it can vary at a different period of time, i.e., the 2nd derivative around the red points is $>$ or < 0. Accordingly, utilizing the 2nd derivative which measures how the 1st gradient of the activations of the neurons are changing, is able to capture the temporal behaviors of the neurons (as in deriving the acceleration from speed) which reflects the curvatures of learned representations.

Due to the VI, the latent space Z hides many different representations that are generated to regularize the VAE; such representations assist in building ESS attention utilizing the behavior of the neurons' activations. To build a visual attention map, our Mgrad$_2$VAE aggregates all multiscale derivatives of the gradient of the latent layer Z concerning each encoding layer, which represents a different scale of dimensionality. For a better visual explanation, our proposed attention map is enforced to follow the original mask distribution, by minimizing the reconstruction and KL losses between the reconstructed mask, attention map, and original mask, simultaneously.

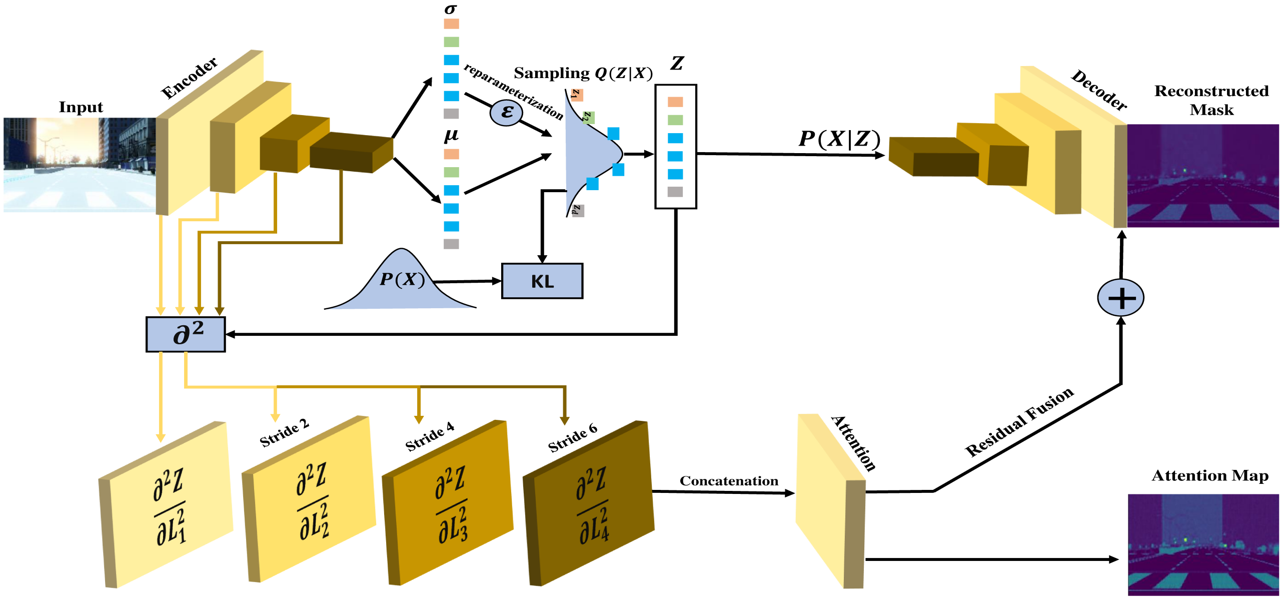

Fig. 2: The Mgrad$_2$VAE block diagram.

3. Mgrad$_2$VAE

Fig. 2 shows our proposed Mgrad$_2$VAE, where it encompasses encoder, decoder, and attention modules. Jointly, the encoder and the decoder include three stages of down-scaling (convolutional neurons with a stride of 2) and up-scaling (de-convolutional neurons with a stride of 2), respectively. Moreover, the Mgrad$_2$VAE visually explains the learned representations utilizing 2^{nd} gradient attention at each encoding scale, i.e., for each encoder's layer there will be a corresponding visual attention map that reflects explainability at that layer, and each attention is enforced to follow the mask distribution to help to contract the representations to the mapped mask.

Moreover, for each layer, the tensor that holds all partial derivatives of the gradient is re-scaled for sake of optimization to match the mask size. Thereafter, all attention maps are aggregated and fused with the $L_{d_{n-1}}$ layer (d_n is the total number of the decoder's layers); such a combination is considered as a novel form of the residual learning [27], which enforces the Mgrad$_2$VAE to learn the residual of mapping between the images and masks by using the 2^{nd} gradient attention. Consequently, besides the explainability of the Mgrad$_2$VAE, it also assists in mask reconstruction by employing the curvatures of activations that are fused to the decoder. Accordingly, the Mgrad$_2$VAE optimizes two losses by using Adam [23] as:

$$\mathbf{L}_{\text{Mgrad}_2\text{VAE}} = \min[\mathbf{L}_{\text{VAE}} + \|X - \theta_{Mgrad}(Z, L_{e_i})\|^2_{\text{Er}}] \qquad (3)$$

where the first loss is obtained from the vanilla VAE [11] that is described at Eqn. (2), and the second loss is the reconstruction loss between the original mask and the aggregated attention at the attention module (see Fig. 2). Furthermore, θ_{Mgrad} reflects the 2^{nd} derivative parameters between the latent space Z concerning all encoder layers L_{e_i}, i.e., for each layer, there will be a corresponding tensor of the size of that layer to allocate all partial derivatives, and the final tensor holds the multiscale attention. Additionally, the model is trained to minimize the loss between each mapped image and its corresponding segmentation mask, and it also optimizes the loss between each attention map that is obtained at a different scale with the same mask; such an optimization enforces all encoder layers to contract to the same data, and it compensates the encoding loss that is raised from down-scaling the dimensionality in the depth layers.

4. EXPERIMENTAL RESULTS

To show the performance of our proposed Mgrad$_2$VAE, we used a collection of SYNTHIA [28] and A2D2 [29] datasets. Specifically, 5600 samples are categorized to the corresponding semantic classes that have been employed. Moreover, the dataset partition to the training and testing subsets complies with $75:25$ protocol, i.e., 75% and 25% of the original data size are the training and testing subsets, respectively. For the sake of computation, in the qualitative analysis, the Mgrad$_2$VAE considers an input layer of the size of $128 \times 256 \times 3$, where the output layer of the size of $128 \times 256 \times 1$. For all experimental works, we consider a minibatch size of 16, and 600 epochs with a learning rate $\eta = 0.001$, where the η is decreased every 100 epoch by a factor of 10^{-2}.

4.1. Qualitative analysis

Our Mgrad$_2$VAE visually explains the learned representations at the neurons activations level through the attention mapping, where it considers the 1^{st} derivative of the gradient (i.e., the 2^{nd} order derivative of neurons activations) between the latent space Z and the encoder layers. For each encoding layer, it produces a tensor to allocate all partial derivatives, and the final attention map can be obtained by concatenating and aggregating (see Fig. 2) all corresponding tensors by using different methods including mean, addition, convolution, etc. Fig. 3 shows the corresponding tensor unfolding (of an image from SYNTHIA dataset) of the attention that is obtained from the last encoding layer L_{e_4}, which represents the last encoding scale as a function of 16 filters depth.

Furthermore, Fig. 4 depicts the final aggregated attention of all encoding layers, where it shows how our model can visually explain the global characteristics of the learned representations at an early stage (L_{e_1}). Moreover, it is also able to show the local characteristics among representations that are captured from the fine-grained features in the depth layers (L_{e_4}).

Fig. 5 shows different examples from the SYNTHIA testing set, ground-truth (GT) masks, reconstructed masks, and the attention maps obtained from our Mgrad$_2$VAE. As it can be noticed from Fig. 4 and Fig. 5, all attention maps which are obtained by our Mgrad$_2$VAE are contracted to the ground truth mask distribution (target domain), and they jointly utilize the multiscale attention mapping (attention at each layer) to build a complimentary map for a

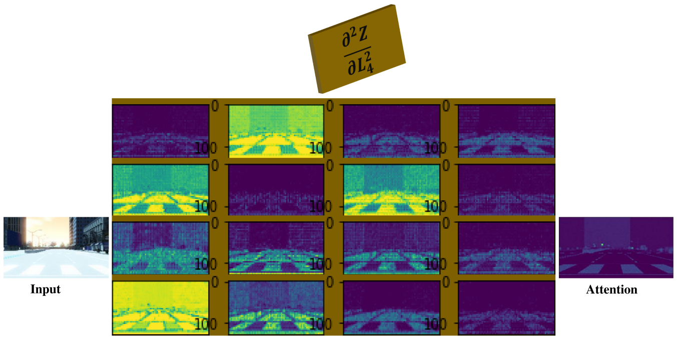

Fig. 3: The 2^{nd} order derivative unfolding of Z with respect to L_{e_4}.

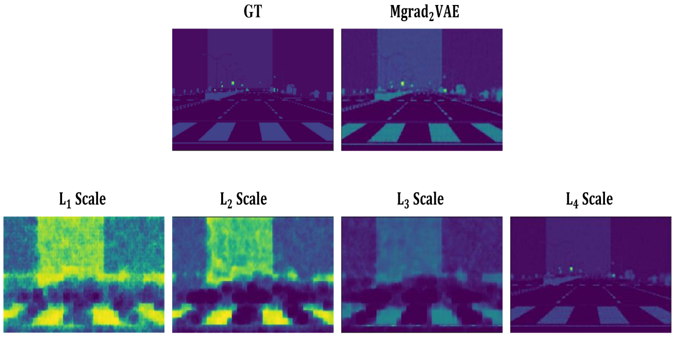

Fig. 4: The multiscale attention of our proposed Mgrad_2VAE, where GT represents the ground truth mask.

better visual explainability.

To assess the semantic structure qualitatively, we employ the SSIM index [30] for both datasets. Moreover, we report the semantic similarities between the ground-truth masks, the corresponding reconstructed masks, and the attention maps in Table 1.

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

Table 1: SSIM of the reconstructed masks and the attentions.

As it can be noticed from Table 1, the Mgrad_2VAE produces an attention map that preserves a similar SSIM index for the reconstructed mask by the decoder, which confirms our methodology and reflects the high quality of the produced attentions.

4.2. Quantitative analysis

Table 2 reports the pixel-wise predictive performance of our proposed Mgrad_2VAE, where we consider the average area under the receiver operator characteristic curve (AUC-ROC) index which reflects an aggregated measure of each pixel classification accuracy. Moreover, we consider the same experimental setup that is reported in section 4. For sake of numerical stability, the depth of the output layer has been adapted from $128 \times 256 \times 1$ to $128 \times 256 \times 3$.

As it can be observed from Table 2, our proposed model offers high performance at the pixel-level classification for both the reconstructed masks and attention maps. Moreover, our attention mapping method outperforms the reconstruction obtained from the decoder side in terms of pixel-level classification in the SYNTHIA dataset.

4.3. Recent work comparison

In this section, we compare our proposed Mgrad_2VAE model with the recent deep learning models, where we consider the deep VAE [11], and the Xception model [31] that has been built based on the U-net architecture [13] and trained on ImageNet dataset [32]. Moreover, we summarize the AUC-ROC metric between the GT masks and the reconstructed masks among all models in Table 3.

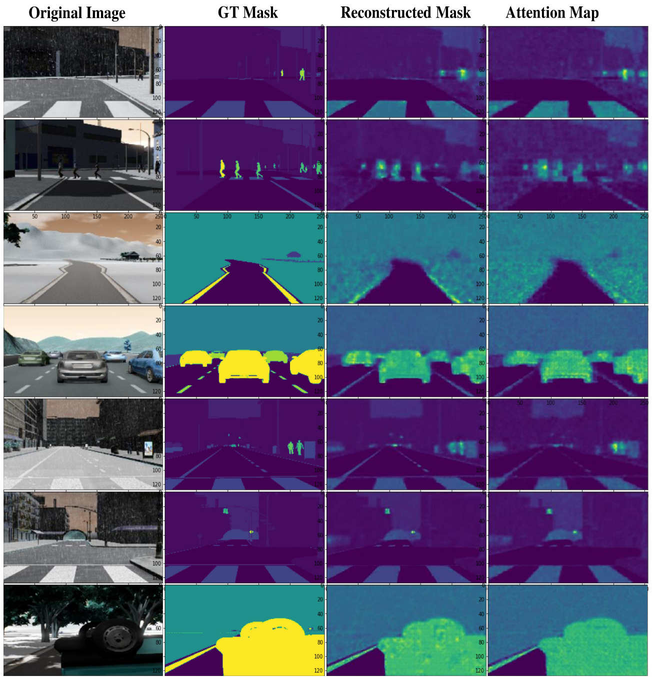

Fig. 5: Examples from the SYNTHIA testing set, where the original images, GT masks, reconstructed masks, and the Mgrad_2VAE attention maps are illustrated from left to right, respectively.

AUC-ROC	SYNTHIA	A2D2
Reconstructed masks	81.50%	95.44%
Attention maps	83.20%	95.36%

Table 2: AUC-ROC of the reconstructed masks and the attentions.

AUC-ROC	SYNTHIA	A2D2
Deep VAE [11]	79.60%	94.05%
Xception [31]	67.43%	95.19%
Our Mgrad_2VAE reconstruction	**81.50%**	**95.44%**
Our Mgrad_2VAE attention	**83.20%**	**95.36%**

Table 3: AUC-ROC comparison with recent deep models.

As it can be seen from Table 3, our proposed Mgrad_2VAE model outperforms all other models in reconstructing masks and attentions. Moreover, although the reconstruction module of our model is typical to the Deep VAE [11], the reconstruction performance of the Mgrad_2VAE is better than [11] by 1.90% and 1.39% for the SYNTHIA and A2D2 datasets, respectively, because of the residual fusion between the decoder and attention modules of our model.

5. CONCLUSIONS

We proposed an explainable VAE model termed as the (Mgrad_2VAE) to be utilized for XAI and EADS applications. Our model uses the multiscale second-order derivative of the neurons' activations of the latent space concerning all other encoding layers. Moreover, it captures the curvature of the learned representations to offer a better visual explainability of the VAE's behavior through attention mapping. Our proposed model outperforms all related deep segmentation models in the quantitative analysis. In future works, we plan to investigate the XAI in harsh environments and rough weather conditions, where the ambient includes rain, snow, dust, fog, etc.

6. REFERENCES

[1] Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, "A survey on unsupervised generative models for exploratory data analysis and representation learning," *Acm computing surveys (csur)*, 2021.

[2] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[4] David Gunning and David Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[5] Junqing Wei, Jarrod M Snider, Junsung Kim, John M Dolan, Raj Rajkumar, and Bakhtiar Litkouhi, "Towards a viable autonomous driving research platform," in *Proc of Intelligent Vehicles Symposium (IV)*, 2013.

[6] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. of ECCV*, 2015.

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun, "CARLA: An open urban driving simulator," in *Proc. of robot learning*, 2017.

[8] A. Genovese, V. Piuri, F. Rundo, F. Scotti, and C. Spampinato, "Pedestrian/cyclist distance estimation from a single rgb image: A cnn-based semantic segmentation approach," in *Proc. of Industrial Technology (ICIT 2021)*, 2021.

[9] Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, "On approximating the nonnegative rank: Applications to image reduction," in *Proc. of CIVEMSA*, 2020.

[10] Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, "Unsupervised learning from limited available data by $\beta-\mathrm{NMF}$ and dual autoencoder," in *Proc. of ICIP*, 2020.

[11] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *Proc. of ICLR*, 2014.

[12] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of Medical image computing and computer-assisted intervention*, 2015.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[15] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus, "Variational ae for end-to-end control of autonomous driving with novelty detection and training de-biasing," in *Proc. of IROS*. IEEE, 2018.

[16] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller, "M2p3: multimodal multi-pedestrian path prediction by self-driving cars with egocentric vision," in *Proc. of ACM Symposium on Applied Computing*, 2020.

[17] Xinyu Chen, Jiajie Xu, Rui Zhou, Wei Chen, Junhua Fang, and Chengfei Liu, "Trajvae: A variational autoencoder model for trajectory generation," *Neurocomputing*, vol. 428, pp. 332–339, 2021.

[18] Andrea Stocco, Michael Weiss, Marco Calzana, and Paolo Tonella, "Misbehaviour prediction for autonomous driving systems," in *Proc. of ICSE*, 2020.

[19] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps, "Towards visually explaining variational autoencoders," in *Proc. of CVPR*, 2020.

[20] Pierre Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. of Unsupervised and Transfer Learning workshop*, 2012.

[21] Geoffrey E. Hinton and Ruslan R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. of ICML*, 2014.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. of ICML*, 2014.

[24] William F Ames, *Numerical methods for partial differential equations*, Academic press, 2014.

[25] Kai Fan, Ziteng Wang, Jeff Beck, James Kwok, and Katherine Heller, "Fast second order stochastic backpropagation for variational inference," in *Proc. of NIPS*, 2015.

[26] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of ICML*, 2010.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016.

[28] Javad Zolfaghari Bengar, Abel Gonzalez-Garcia, Gabriel Villalonga, Bogdan Raducanu, Hamed Habibi Aghdam, Mikhail Mozerov, Antonio M Lopez, and Joost van de Weijer, "Temporal coherence for active learning in videos," in *Proc. of ICCVW*, 2019.

[29] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Muhlegg, Sebastian Dorn, et al., "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:2004.06320*, 2020.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[31] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of CVPR*, 2017.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. of NIPS*, 2012.