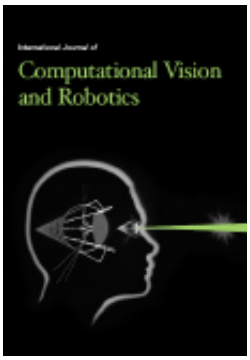


International Journal of Computational Vision and Robotics

ISSN online: 1752-914X - ISSN print: 1752-9131
https://www.inderscience.com/ijcvr

Unsupervised image transformation for long wave infrared and visual image matching using two channel convolutional autoencoder network

Kavitha Kuppala, Sandhya Banda, S. Sagar Imambi

DOI: 10.1504/IJCVR.2022.10050246

Article History:

Received:	23 September 2021
Accepted:	24 June 2022
Published online:	01 December 2023

Unsupervised image transformation for long wave infrared and visual image matching using two channel convolutional autoencoder network

Kavitha Kuppala*

Department of Computer Science and Engineering,
K.L. University,
Guntur, Andhra Pradesh, India
Email: kavithakbjr@gmail.com
*Corresponding author

Sandhya Banda

Department of Computer Science and Engineering,
Maturi Venkata Subba Rao Engineering College,
Hyderabad, Telangana, India
Email: sandhyab16@gmail.com

S. Sagar Imambi

Department of Computer Science and Engineering,
K.L. University,
Guntur, Andhra Pradesh, India
Email: simambi@gmail.com

Abstract: Pixel level matching of multi-spectral images is an important precursor to a wide range of applications. An efficient feature representation which can address the inherent dissimilar characteristics of acquisition by the respective sensors is essential for finding similarity between visual and thermal image regions. Lack of sufficient benchmark datasets of corresponding visual and LWIR images hinders the training of supervised learning approaches, such as CNN. To address both the issues of nonlinear variations and unavailability of huge data, we propose a novel two channel non-weight sharing convolutional autoencoder architecture, which computes similarity using encodings of the image regions. One channel is used to generate an efficient representation of the visible image patch, whereas the second channel is used to transform an infrared patch to a corresponding visual region using encoded representation. Results are shown by computing patch similarity using representations generated from various encoder architectures, evaluated on two datasets.

Keywords: convolutional autoencoder; CAE; multi-spectral image matching; transformation network; two channel siamese architecture; structual similarity measure; SSIM; KAIST dataset; mean squared error; MSE; peak signal to noise ratio; PSNR; Earth mover's distance; EMD.

Reference to this paper should be made as follows: Kuppala, K., Banda, S. and Imambi, S.S. (2024) 'Unsupervised image transformation for long wave infrared and visual image matching using two channel convolutional autoencoder network', *Int. J. Computational Vision and Robotics*, Vol. 14, No. 1, pp.63–83.

Biographical notes: Kavitha Kuppala is currently pursuing her PhD in Computer Science and Engineering from the K.L. University, Vijayawada, India, she also completed her graduation in Computer Science in 2002 from the St. Joseph's Degree College, Kurnool, affiliated to Sri Krishnadevaraya University. She possesses a Master's in Computer Applications from Sri Padmavati Mahila Visva Vidyalayam, Tirupati. She also has to her credit an MTech in Computer Science and Engineering in 2010 from the Aurora's Technological and Research Institute Hyderabad, affiliated to Jawaharlal Nehru Technological University, Hyderabad. She has been serving as a faculty in the CSE and IT Departments at Aurora's Technological and Research Institute, Hyderabad since 2006. Her areas of interest include data mining and deep learning.

Sandhya Banda completed her graduation in 1999 from the Osmania University in Electrical and Electronics Engineering. She has completed her Master's in Software Systems from the BITS Pilani in 2005. She was awarded PhD in Computer Science from the University of Hyderabad in 2011. Her professional experience is a rich blend of industry, R&D and academic spanning more than 16 years. Her principal areas of research include machine learning, deep learning, and image processing. She is currently serving as a Professor in the Department of Computer Science and Engineering in MVSR Engineering College, Hyderabad.

S. Sagar Imambi is an Professor at the Department of Computer Science and Engineering at the K.L. Educational Foundation, Deemed to be University, Vaddeswaram, and Andhra Pradesh. She is having 22 years of teaching and 11 years of research experience. She has published 37 peer review research articles in various journals, 32 articles at various international conferences. She published book chapters with De Gyter Publications, NewJersy, and Springer. She is also an editorial board member of five academic journals and reviewer for international conferences. She is associate with IEEE and ACM membership. Under her guidance seven scholars awarded MPhil. She is guiding six PhD scholars. Her main areas of research interest are data mining, text mining, machine learning and deep learning.

1 Introduction

Infrared and visual image matching is challenging due to inherent difference in the way images are acquired by the multi-sensors. Conventional approaches of patch matching fail to capture nonlinear variations between the images and hence cannot be adopted for pixel level matching of IR and visible (VS) images. Convolution neural network's (CNN's) have fast emerged as an efficient approach for image feature extraction, to be used across wide range of applications including matching. Several publications

have shown the superiority of features extracted from pre-trained CNN's, such as Krizhevsky et al. (2012), Simonyan and Zisserman (2014) and Szegedy et al. (2015), when compared to handcrafted features. To better address the challenges in image matching such as the wide range of photometric variations and lack of datasets with balanced class distrbution, siamese and triplet CNN's have been proposed (Zagoruyko and Komodakis, 2015; Wang et al., 2014; Balntas et al., 2016; Hani Altwaijry and Belongie, 2016). Networks trained for computing the similarity between patches with similar modality cannot be directly adapted for multi-spectral IR and VS patch matching. Inorder to adapt a transfer learning approach using pretrained CNN networks for image matching, a benchmark dataset of corresponding multi-spectral patches is mandatory. This has motivated us towards adapting an unspervised learning-based autoencoder architectures.

Autoencoders are built using encoder and decoder networks, the functional objective being reconstruction of the compressed feature vector. Such features are superior in comparison with features extracted using traditional approaches (Jaques et al., 2017). Taxonomy of encoders is based on latent space learning and arrangement of layers as shown in Figure 1. In deterministic learning the flow/ communication through the kernel maps is static, whereas in probabilistic approach these maps are randomly propagated across the layers. Other way of differentiating encoders is based on positioning array of layers and connections among them; In symmetric architecture both encoder and decoder networks have identical structure of arranging layers, whereas in asymmetric networks, structure differs between encoder and decoder.

Generally, autoencoders are designed based on the type of data, arrangement of layers and the encoding required. Some of the variants of encoder acting as a baseline for the advanced encoding schemes are denoising autoencoder (DAE), convolutional autoencoder (CAE), contrastive autoencoder (Bank et al., 2020), variational autoencoder (VAE) (Baldi, 2012), stacked autoencoder (SAE) (Benbakreti et al., 2021) and combination can be any of the above. For example DCVA (Zilvan et al., 2019) is denoising convolutional VAE, generative autoencoders (Meng et al., 2017), LSTM encoder (Song et al., 2018), stacked convolutional DAE (Du et al., 2016), etc.

Figure 1 Classification of autoencoders

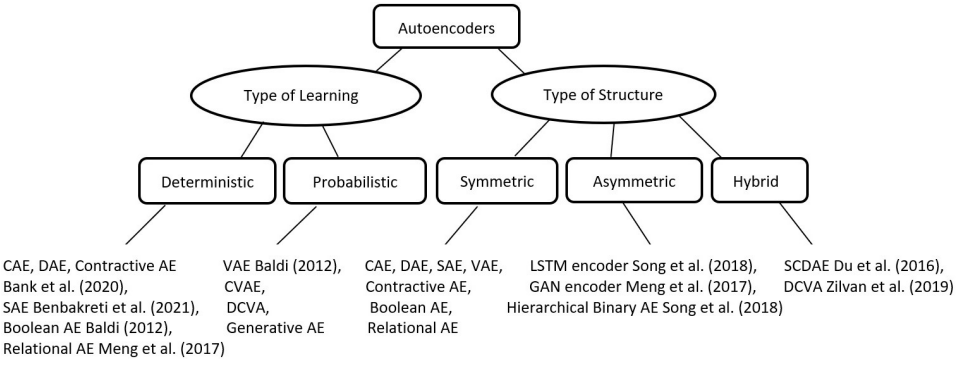


Table 1 Survey of CAEs highlighting network parameters, evaluation metrics methods across applications

	Architecture parameters		*Architecture hyperparameters*			*Initial*				
Study	*Type of images*	*Type of structure and no. of conv layers*	*Pooling sampling*	*Activation function*	*Regulariser, normalisation*	*learning rate*	*Optimiser*	*Loss*	*Evaluation metric*	*Application/s*
Mao et al. (2016)	CT	Symmetric 5conv-5deconv	Upsampling	ReLU	L2	0.0001	Adam	MSE, Euclidean loss	SSIM and PSNR	Image restoration, and super resolution and super resolution
Chen et al. (2017)	Visible	Symmetric 3conv-3deconv	Maxpooling	ReLU Sigmoid		0.0001	SGD	MSE and cross entropy reconstruction loss	Precision recall, F1, AUC, ROC and accuracy	Classification
Knyaz et al. (2017)	Thermal	Symmetric 2conv-2deconv	Pooling	Sigmoid		0.001	SGD	Euclidean cross entropy loss	Precision recall and AUC	3D object and image reconstruction
Van Tulder and de Bruijne (2018)	MRI	Asymmetric 4conv-4deconv with dense layer		LeakyReLU	BN and D (dropout)	0.3	SGD	SSIM combined reconstruction loss	MI, SD and NCC	Segmentation and classification
Berg et al. (2018)	LWIR-RGB	Asymmetric 4conv-3deconv	Upsampling	LeakyReLU	BN and D	0.001	Adam	SSIM	SSIM, PSNR, L1 and RMSE	Transformation
Nyberg et al. (2018)	TIR-RGB	Asymmetric 4conv-3deconv	Upsampling	LeakyReLU	BN and D	0.001	Adam	SSIM	SSIM, PSNR, L1 and RMSE	Transformation
Li and Wu (2018)	Infrared-RGB	Asymmetric 1conv with 3 dense layers-4deconv	Maxpooling and upsmapling	ReLU	BN	0.0001	RmsProp	SSIM, L1 and average pixel loss	SSIM	Fusion
Ali et al. (2019)	MRI	Asymmetric 6conv-5deconv	Maxpooling and upsmapling	ReLU ReLU	DA, BN and L2	0.0001 0.0001	Adam Adam	MSE and binary cross entropy loss	Accuracy	Classification
Laakom et al. (2019)	RGB	Symmetric 4conv-4deconv	Maxpooling	ReLU	DA and D	0.01	Sigmoid	Composite loss	Recovery angular error	Colour constancy
Kerner et al. (2019)	RGB	Symmetric 3conv-3deconv	Upsampling	ReLU	BN and D	0.0001	RMSProp	MSE loss	AUC and ROC	Classification and detection
Cheng et al. (2018)	RGB	Symmetric 6conv-6deconv	Down-upsampling	PReLU	BN	0.0001	Adam	MSE loss	SSIM and PSNR	Image compression and reconstruction
Žižakić et al. (2019)	RGB	Symmetric 3conv-3deconv	Maxpooling and upsampling	ReLU		0.000001	Adadelta	Binary cross entropy	Sum of squared differences	Image inpainting

Notes: D – dropout, BN – batch normalisation, DA – data augmentation, MI – mutual information, MSE – mean squared error, NCC – normalised cross correlation, AUC – area under curve and SGD – stochastic gradient descent.

Table 1 Survey of CAEs highlighting network parameters, evaluation metrics methods across applications (continued)

	Architecture parameters		*Architecture hyperparameters*			*Initial*				
Study	*Type of images*	*Type of structure and no. of conv layers*	*Pooling sampling*	*Activation function*	*Regulariser, normalisation*	*learning rate*	*Optimiser*	*Loss*	*Evaluation metric*	*Application/s*
Luppino et al. (2020)	NIR-RGB	Symmetric 4conv-4deconv		LeakyReLU and tanh	D	0.0001	Adam	Code correlation, weighted translation, cycle-consistency and reconstruction loss	Accuracy, true positive, negatives and Cohen's kappa coefficient	Change detection
Arif and Mahalanobis (2020)	Midwave infrared Midwave infrared	Symmetric 4conv-4deconv	Pooling and upsampling	LeakyReLU		0.0001	Adam	Reconstruction loss, MSE	Accuracy	View generation and classification
Kolberg et al. (2020)	RGB, SWIR and NIR	Symmetric 1conv-1deconv	Maxpooling and upsampling	ReLU	D	0.001	RMSProp	MSE	AUC	Anomaly detection
Kerner et al. (2020)	Multi-spectral satellite images	Symmetric 3conv-3deconv		LeakyReLU	DA	0.001	Adam	SSIM and MSE	L2 and SSIM	Planetary exploration

Notes: D – dropout, BN – batch normalisation, DA – data augmentation, MI – mutual information, MSE – mean squared error, NCC – normalised cross correlation, AUC – area under curve and SGD – stochastic gradient descent.

Despite the main goal, i.e., compression and reconstruction, encoders are being modelled to fill the gap of efficient image representation across applications. Zhang (2018) proved that convolutional encoding (CAE) is better in capturing and preserving the spatial information of images because of convolution layers. In Kerner et al. (2019), the results presented for image classification show that detecting and selecting novel features by using CAE out-perform the various classical supervised training methodologies in spectrally diverse images. When the CAE model weights are loaded/initialised to CNN kernels, the classification results surpass those obtained with pretrained CNN (Masci et al., 2011).

We propose unsupervised 2-channel CAEs trained with different spectral images to overcome the illuminational variance. Our major contributions are the following:

1 Use of unsupervised learning-based CAE for matching long wave IR and VS images.

2 Use of transformation CAE in a siamese architecture for image matching application.

The paper is organised as follows. In Section 2 survey about CAE's across various applications is presented, Section 3 the image similarity between multi-spectral images with need of colour transformation and 2-channel network is described. In Section 4 preparation of data, architectural design and fine-tuning with various parameters is provided while results are presented in Section 5. The paper is concluded in Section 6.

2 Background

The survey focuses on use of CAE across applications with varying architectures and loss functions. Dong et al. (2018) review generic autoencoder model exploring kinds of autoencoders against the hyperparameters of the models. The study demonstrated that CAEs are relatively more stable than CNNs in recognition-based applications.

Survey by Georgiou et al. (2020), explores various methodologies ranging from conventional non-deep to contemporary deep learning using datasets with varying dimensionality and modality. Among all the invariant features, CAE model is found effective for object classification and anamoly detection by Pawar and Attar (2020).

In a CAE the encoder performs convolution, while the decoder is responsible for deconvolution and up-sampling. The key idea is to apply the up-sampling decoder network, which maps the low-resolution encoder feature maps. This architecture substantially reduces the number of trainable parameters and reuses the encoder's pooling indices to up-sample. Table 1 lists the papers in which CAE is used for various applications with focus on architecture, hyperparameters and evaluation metrics.

2.1 Architecture of CAE: layers and hyperparameters

In Masci et al. (2011), a ladder is built with CAE kernels, to extract unsupervised features. CAE kernel maps/features points which are preserved consistently without applying any regularisation, are assigned for classification. This approach proved superior in comparison with both the basic autoencoder and CNN. Similar idea is used by Tharani et al. (2018) in classifying remote sensing images, where in a deep residual

encoder-decoder network is trained for feature extratction along with a discrimination network. Shahriari (2016) added Fisher linear discriminant as a functional layer for improvement in retrieving texture feature identification.

Volodymyr et al. adopted a CAE and experimented with presence of pool and unpool layers for classification application. Results prove that the architecture with pool and unpool layers is capturing and storing the boundary information in the encoder feature maps (Turchenko et al., 2017).

Yasrab et al. (2017) designed encoder layers similar to VGG16 for driver assistance system. The authors conducted tests to identify the best possible architecture and set of hyperparameters. It is found that by using dropout in encoder, the model is less sensitive to the trained images and generalised better. Among the activation functions, with ELU the pixel values reach zero mean which helps in learning good representatives. Addition of normalisation layer reduces the complexity range of pixel values in feature maps with increased accuracy.

Clevert et al. (2015) experimented various activation functions for CAEs in classifying CIFAR-10 and 100. Results reported that ELU outperforms other activation functions such as ReLU, LeakyReLU, PReLU and RReLU. Van Tulder and de Bruijne (2018) used CAE for cross modal classification.

To address differences between modalities, an axial neural network architecture with CNNs is used with a separate network path for each input source (Van Tulder and de Bruijne, 2018).

Masanori et al. used evolutionary algorithm to find the best CAE that fits for the inpainting and restoration problems. The algorithm upon successive search, identified a simple CAE with ADAM and L2 loss best. Results as prove that this CAE is good in comparison with complex generative models (Suganuma et al., 2018).

2.2 Training loss and data augmentation

Sadegh et al. adopted SegNet-based with encoder-decoder network to correctly discriminate the rock images that have colour and texture variance. Data augmentation is applied to generate the synthetic data for training, to increase the performance (Karimpouli and Tahmasebi, 2019). Cheng et al. (2018) stated that the CAE further code compressed with PCA, quantisation and entropy layers is powerful in attaining high coding efficiency in image compression. Rate distorted loss is used to optimise CAE.

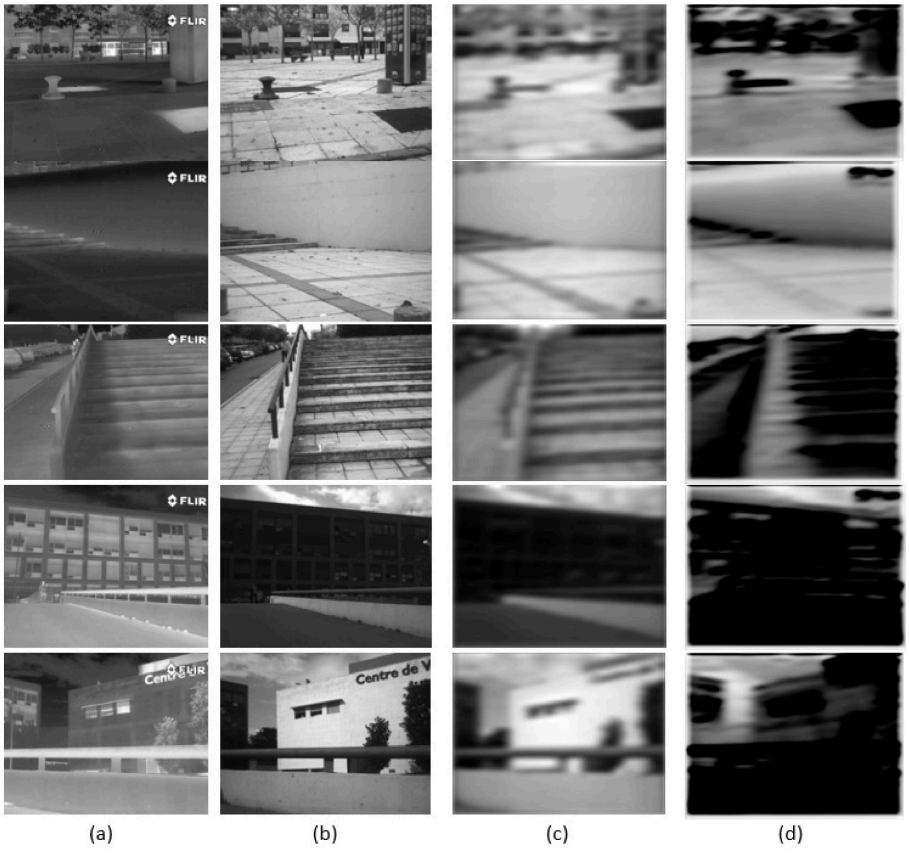
Azarang et al. (2019) trained CAE with a hybrid loss constructed using MSE and SSIM for image fusion. In Guo et al. (2017), CAE is used in clustering application by simultaneously minimising the reconstruction loss of CAEs and the clustering loss.

3 Multi-spectral image similarity using CAE

A 2-channel/Siamese frameworks perform well for image matching and similarity applications (Zagoruyko and Komodakis, 2015; Wang et al., 2014). Appalaraju and Chaoji (2017) proved that multi-scale Siamese CNN is better for finding fine-grained features compared with CNN in curriculum learning. Later in Liu et al. (2018) similar framework is adopted in H-Net for cross-domain image matching with MSE as a function. H-Net has given prominent results in matching but failed in retrieval. Liu et al.

(2018) extended the work and built HNet++, with MSE and contrastive loss. Features extracted from the model are robust and invariant.

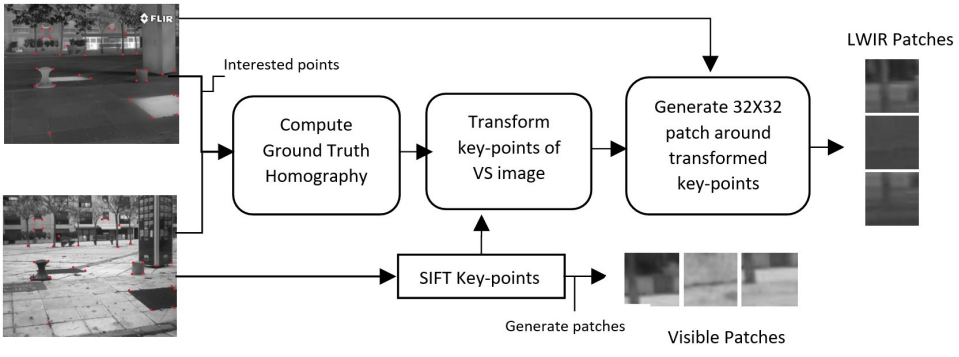
Figure 2 Sample images, (a) original LWIR (b) original visible images (c) unimodal-net VS encoded representative images (d) encoded represenative images from multi-modal-net with SSIM loss



Results observed are promising in both cross-domain patch matching and retrieval. Both H-Net and HNet++ have shown the efficacy of the network using sythetic images generated by cycleGan consisting of illumination variation. However, finding similarity between long wave infrared and visible images is much more challenging because of the wide disparity in the wavelengths of visible [8 μm to 15 μm] and infrared [0.4 μm to 0.7 μm]. To overcome the problem of LWIR and VS image matching, the study of image transformation is conducted. Visible spectrum images are widely used, but most of the information will be lost or not visible with loss of light which can be captured with thermal infrared imaging. To exploit the advantages of both visible and thermal images, most of the multi-modal imaging applications are applying image transformation techniques. We present some of the papers which concentrate on visible and thermal image transformation using CAE architectures. Berg et al. (2018) used encoder-decoder which mainly consists of convolution, batch normalisation, LeakyReLU, drop out for transforming thermal to visible images. Results were analysed quantitatively with L1, RMSE, PSNR and SSIM which show that

estimating illuminance and chrominance separately resulted in better outcome. Nyberg et al. (2018) trained auto-encoder architecture using different colour spaces and applied on cross-spectral image transformation between infrared and visible images. Laakom et al. (2019) composite loss built with binary cross entropy and recovery angular loss is used in colour constancy problem. Results are better and much generalised with fewer parameters.

Figure 3 Generating groundtruth from SIFT keypoints



4 Proposed network architecture

Objective of the proposed work is to find similarity between patches (/regions) of visible and long wave infrared (LWIR) images. Though the problem of multi-spectral image matching has been addressed in the literature to a certain extent, finding similarity between LWIR and visible is still challenging. The problem is addressed using a 2-channel CAE architecture, designed as follows.

a One channel is a CAE, which is trained to encode unimodal VS or IR image.

b Second channel is a CAE, designed as a multi-modal IR to VS 'transformation' network.

Both the channels of the architecture are trained independently with patches of LWIR and Visible images. The encoder part of both the CAEs is combined and lambda layer is added to compute similarity using SSIM.

4.1 Unimodal IR, VS single channel encoder

The unsupervised CAE methods, that we focus here are primarily intended to extract the representations that can serve as features for finding similarity. First, we train the unimodal single channel VS encoder with visible patches and next, IR encoder with IR image patches as shown in Figure 4(a).

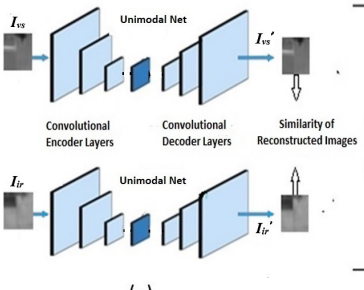
We can define the E: encoding and D: decoding functions as F and G,

$$F : E(I_{ir}) \rightarrow I_{ir}^{U}; G : D(I_{ir}^{U}) \rightarrow I'_{ir} \tag{1}$$

$$F : E(I_{vs}) \rightarrow I_{vs}^{U}; G : D(I_{vs}^{U}) \rightarrow I'_{vs} \tag{2}$$

where I_{ir}, I_{vs} represent IR patch, VS patch respectively; I_{ir}^{U}, I_{vs}^{U} for coded feature representative and I'_{ir}, I'_{vs} are patches obtained by decoding the coded features.

Figure 4 (a) 2-channel unimodal CAE (b) Layer wise details of the model (see online version for colours)



(a)

	Layers	Size x kernels		Layers	Size x kernels
Encoder	input	32x32x1	Decoder	Upsampling	8x8x16
	Conv+ BN+ ELU	32x32x96		Conv_T+ BN+ ELU	8x8x32
	Maxpooling	16x16x96		Upsampling	16x16x32
	Conv+ BN+ ELU	16x16x64		Conv_T+ BN+ ELU	16x16x64
	Maxpooling	8x8x64		Upsampling	32x32x64
	Conv+ BN+ ELU	8x8x32		Conv_T+ BN+ ELU	32x32x96
	Maxpooling	4x4x32		Conv_T+ BN+ ELU	32x32x1
	Conv+ BN+ ELU	4x4x16		Output	32x32x1

(b)

Table 2 Proposed network and training parameters

Type of DNN	*CAE*
Optimiser	RMSProp
Activation function	ELU
Conv and deconv layers	4conv-4deconv
Pooling and upsampling	Maxpooling and upsampling
Regulariser	BN
Learning rate	0.0001
Loss	MSE and SSIM

We used a dataset of 100 pairs of visible and long wave infrared images of size 506 × 506 (Campo et al., 2012). Dataset is freely available at http://www.cvc.uab.es/adas/projects/simeve. Figures 2(a) and 2(b) show some of the sample LWIR and visble images from the dataset.

We started our experimentation by training a CAE with a VS image of size 400 × 400. The encoded representatives of CAEs are presented in Figure 2(c). It is clearly visible from the obtained images, that this approach fails to preserve the pixel level information. To overcome such problems, we restated our objective to find patch level similarity.

Data generation: we extracted overlapping patches of size 32 × 32 from each image of size 506 × 506. A total of 178,600 patches are extracted from 100 images, out of which 50% are used for training and 25% for validation.

The unimodal encoder encompasses convolutional, max pooling and batch normalisation layers. Details are as follows: C(96,3,1)-BN-ELU-P(3,2)-C(64,3,1)-BN-ELU-P(3,2)-C(32,3,1)-BN-ELU-P(3,2)-C(16,3,1)-BN-ELU-P(3,2) shown in Figure 4(b), representation is similar to Liu et al. (2018): $C(n, k, s)$ is the convolution with n filters of kernel size $k \times k$, $P(k, s)$ the max pooling of size $k \times k$ and with the stride s.

Figure 5 2-channel CAEs with transformer in multi-modal-net (see online version for colours)

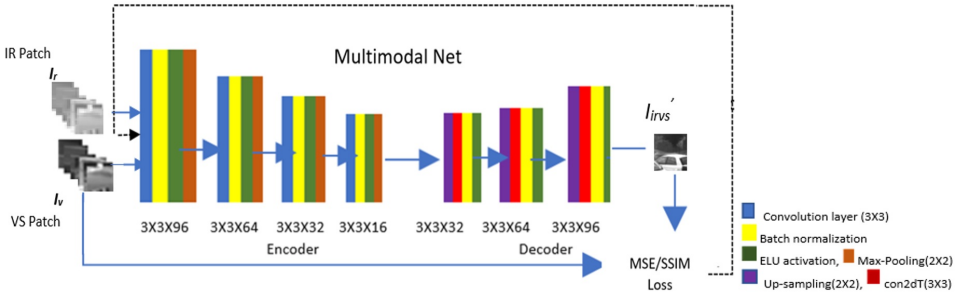


Figure 6 (a) LWIR patches (b) Visible patches of size 32 × 32 from image (c) Unimodal-net IR decoded patches (d) Unimodal-net VS decoded patches (e) Multi-modal-net IR to VS with MSE as loss (f) Multi-modal-net IR to VS with SSIM as loss

For the decoder, transpose convolution layer is used to reconstruct patches of size 32 × 32 from encoded feature space. The details of decoder architecture are: TC(16,3,1)-BN-ELU-P(3,2)-TC(32,3,1)-BN-ELU-P(3,2)-TC(64,3,1)-BN-ELU-P(3,2)-TC(96,3,1)-BN-ELU-P(3,2)-Sigmoid. TC (n, k, s) is a transposed convolution with n filters of size $k \times k$ applied with the stride s. Feature space/representation is compressed by decreasing the number of kernels at each convolutional layer of encoder and increase of kernels at decoder to reconstruct the same image. Network is trained with a minibatch size of 50 and learning rate of 0.0001. MSE is used as a loss function, defined as follows for IR images and the same can be used with VS:

$$MSE(I_{ir}, I'_{ir}) = \| I_{ir} - I'_{ir} \|^2 \tag{3}$$

4.2 Multi-modal IR to VS encoder

To overcome the challenge of extreme illumination variations, a network is trained to generate a representation of LWIR image patch which when decoded results in a corresponding visble patch.

We trained 2-channel CAE with a pair of corresponding IR and VS images of same size. The encoded representatives of CAEs are presented in Figure 2(d). The obtained visible images after transformation, lose the pixel level information. To overcome such problems, we experimented with patches.

Data generation: to train the multi-modal encoder corresponding sets of IR-VS image patches need to be genarated. Figure 3, depicts the patch extraction process and details are given below:

1 Compute ground truth homography between LWIR and visible images, by manually identifying five corresponding pixels between the images.

2 Compute keypoints of VS image using SIFT (Lowe, 2004) as detector.

3 Transform the keypoints of visible image to IR image using ground truth homography.

4 A patch of size 32 × 32 is extracted around each keypoint in both LWIR and visible images.

A total of around, 3,000 patches are extracted from each image and sum of 300,000 patches are generated for 100 images.

The multi-modal-net encoding and decoding functions are as follows,

$$F : E(I_{ir}, I_{vs}) \rightarrow I^M_{ir}; G : D(I^M_{ir}) \rightarrow I'_{irvs}, \tag{4}$$

where I'_{irvs} is patch decoded from I^M_{ir} representative code of multi-modal IR and VS patches.

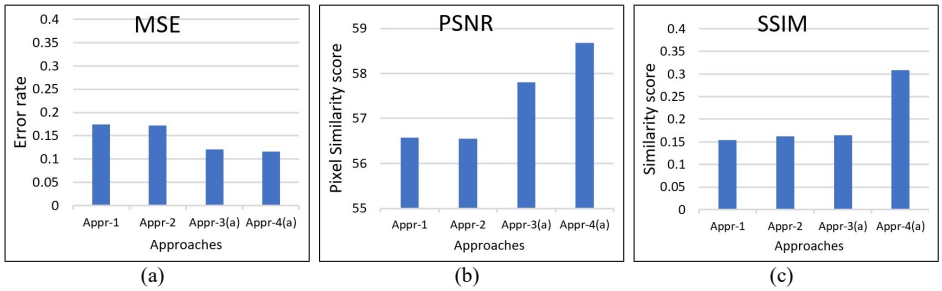
The idea is to use CAE as a transformation network with input being corresponding IR and VS patches of size 32 × 32 as shown in Figure 5. The network gets trained by computing loss between encoded IR image patch and VS patch. To attain good quality reconstruction of compressed image, with less structural degradation, structural similarity index (SSIM) is used (Berg et al., 2018; Nyberg et al., 2018).

The structural difference score between the input and output patches is calculated with DSSIM as follows:

$$DSSIM(I'_{irvs}, I_{vs}) = 1/2(1 - SSIM(I_{vs} - I'_{irvs})) \tag{5}$$

Similar batch size and the learning rate are maintained in training, as in unimodal CAE. Batch normalisation (BN) is applied on the each batch of output feature maps from convolution layers. To maintain the similar range of values in pixel intesities along the batch, features are normalised with a batch mean and variance. Small constant values are used to maintain the smoothness, avoid vanishing gradient problem and speedup the training process. BN, max-pooling, up-sampling are applied, inspired by Žižakić et al. (2019). A simplified set of parameters used in building our network are shown in Table 2.

Figure 7 Average of, (a) MSE (b) PSNR (c) SSIM similarity for 16,000 similar patches (see online version for colours)



(a) (b) (c)

Figure 8 Average SSIM score over each image for original VS patches (see online version for colours)

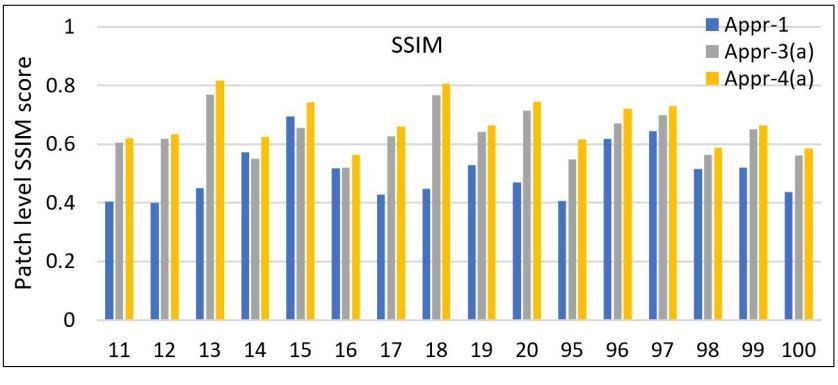
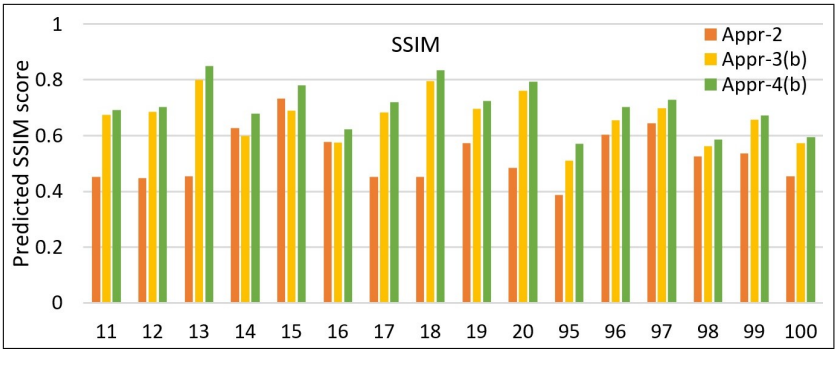


Figure 9 Average SSIM score over each image for representative encoded VS patches (see online version for colours)



5 Evaluation and results

Implementation is carried out on Intel Core i3-3220 CPU @ 3.30 GHz, 8 GB RAM, Ubuntu 16.04 64bit GPU: Nvidia GeForce GT 710. OpenCV 3.3 DNN module is used for all the approaches.

Evaluation is carried out in the following ways:

- Approach-1 (Appr-1): Distance between original IR patch (I_{ir}) and VS patch (I_{vs}) in order to provide a baseline.
- Approach-2 (Appr-2): Distance between encoded representative patches of IR (I'_{ir}) and VS (I'_{vs}) from unimodal CAEs.
- Approach-3: Multi-modal network with MSE as loss:
  1 Appr-3.a: Distance between encoded representative (I'_{irvs}) from multi-modal encoder trained with MSE as loss and VS original patch (I_{vs}).
  2 Appr-3.b: Distance between encoded representative (I'_{irvs}) from multi-modal encoder trained with MSE as loss and unimodal-net VS encoder (I'_{vs}).
- Approach-4: Multi-modal network with SSIM as loss:
  1 Appr-4.a: Distance between encoded representative (I'_{irvs}) from multi-modal encoder trained with SSIM as loss and VS original patch (I_{vs}).
  2 Appr-4.b: Distance between encoded representative (I'_{irvs}) from multi-modal encoder trained with SSIM as loss and unimodal-net VS encoder (I'_{vs}).

5.1 Evaluation using IR-VS dataset

In Figure 6, we present some of the randomly picked decoded patches for unimodal and multi-modal networks. The average similarity scores of the decoded patch representives are presented in Table 3.

Table 3 Similarity score of evaluation approaches on IRVS

Patch evaluation	*Avg. SSIM score*	*Decoded patch evaluation*	*Avg. SSIM score*
Appr-1 (I_{ir}, I_{vs})	0.59	Appr-2 (I'_{ir}, I'_{vs})	0.63
Appr-3.a (I'_{irvs}, I_{vs})	0.72	Appr-3.b (I'_{irvs}, I'_{vs})	0.75
Appr-4.a (I'_{irvs}, I_{vs})	*0.79*	Appr-4.b (I'_{irvs}, I'_{vs})	*0.83*

For quantitative evaluation, we used mean squared error (MSE), peak signal to noise ratio (PSNR), and structural similarity (SSIM) as in Berg et al. (2018) and Tsagkatakis et al. (2019). In evaluating the trained model, we have taken 16,000 patches from 16 images, out of which ten images belong to training and validation dataset, and 6 images belong to test set. The mean values of the MSE, PSNR and SSIM evaluation metrics for all the four approaches can be seen in Figures 7(a), 7(b) and 7(c) respectively. The following observations are made:

- The error rate between the original IR patches and VS patches (Appr-1) is more in comparison with any other encoder-based (Appr-2, Appr-3 and Appr-4) approach.
- Performance of multi-modal CAE (Appr-3 and Appr-4) is much better as compared to unimodal CAE (Appr-2).

- Multi-modal CAE network, trained with SSIM (Appr-4) is better in comparison with MSE (Appr-3) trained network.

In Figure 8, we present the graphs which depict structural similarity scores between original VS patches and patches from encoded representatives as mentioned in approaches 1, 3a, 4a for the 16 images. In Figure 9, SSIM scores between VS unimodal encoded representatives and patches of encoded representatives as in 2, 3b, 4b.

It can be observed that the transformation encoder trained with SSIM loss is able to generate visual patches close to the original VS patch.

Table 4 Similarity and dissimilarity scores of evaluation approaches on KAIST

Similar patch evaluation	*RMSE*	*SSIM*	*PSNR*	*EMD*	*Dissimilar patch evaluation*	*RMSE*	*SSIM*	*PSNR*	*EMD*
Appr-1 (I_{ir}, I_{vs})	0.326	0.226	56.65	118.758	Appr-1 (I_{ir}, I_{vs})	0.105	0.233	60.69	29.853
Appr-4.a (I'_{irvs}, I_{vs})	*0.015*	*0.826*	*68.041*	*57.504*	Appr-4.a (I'_{irvs}, I_{vs})	*0.118*	0.515	60.35	*146.907*

Table 5 Precision, recall, F-measure and accuracy table

Threshold measure	*0.8*		*0.7*		*0.6*	
	Appr-1 (I_{ir}, I_{vs})	*Appr-4.a* (I'_{irvs}, I_{vs})	*Appr-1* (I_{ir}, I_{vs})	*Appr-4.a* (I'_{irvs}, I_{vs})	*Appr-1* (I_{ir}, I_{vs})	*Appr-4.a* (I'_{irvs}, I_{vs})
Precision	0.758	0.969	0.655	*0.877*	0.648	*0.828*
Recall	0.066	0.655	0.107	*0.8*	0.179	*0.880*
F-measure	0.115	0.778	0.174	*0.835*	0.263	*0.851*
Accuracy	0.532	0.817	0.522	*0.837*	0.521	*0.844*

5.2 Evaluation using KAIST dataset

We evaluated our model with a popular multi-spectral dataset, KAIST pedestrian dataset (Berg et al., 2018; Nyberg et al., 2018). The dataset has 95 k real traffic scene images, captured both in the day and night time from the moving vehicle with different sensors. Figure 10, shows a sample LWIR image and its corresponding visible image, each of size 640 × 512. The model is trained with approximately 3,000 day time image patches for 50 epochs with SSIM as loss.

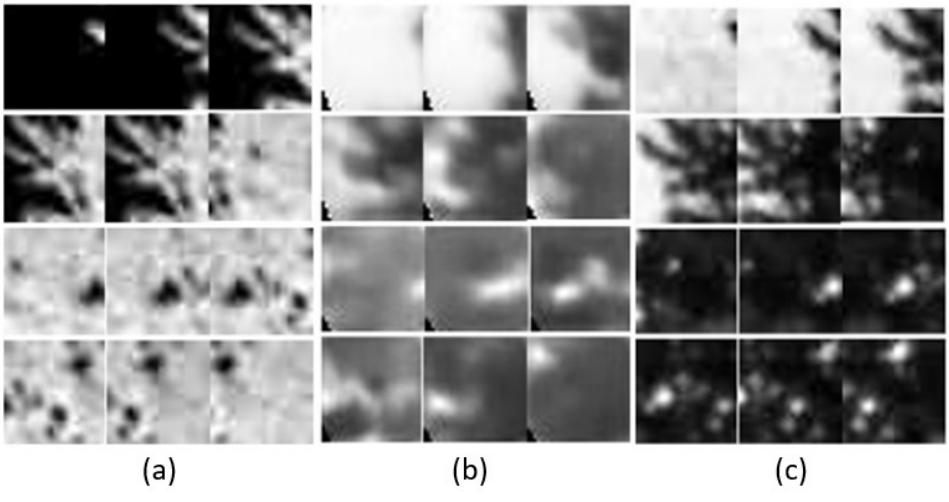
Some of the patches extracted from VS and LWIR images of Figure 10, and their corresponding decoded patches obtained using multi-modal-net are shown in Figure 11.

The network is evaluated for use in image matching, wherein corresponding patches between images must be found using nearest neighbour approach. For a given region/patch of an image, corresponding patch in the other image is found mostly based on a threshold set on the similarity value between the patches. Hence, we analysed the various similarity and dissimilarity measures computed between the encoded representations of the patches. Rubner et al. (2000) and Zhang et al. (2020) stated that Earth mover's distance (EMD) is one of the best metric in the area of image retireval and image matching. It is able to find optimal structural difference in representations of image regions.

Table 4 shows the average values of the measures, computed over 16,000 similar and dissimilar patches of visual and IR images. It can be observed that encoded representations help in clearly distinguishing between positive and negative pairs.

Figure 10 Original IR and visible images (see online version for colours)

Figure 11 (a) IRpatches (b) Decoded patches from multi-modal-net (c) VSpatches



Nyberg et al. (2018) proposed (TIR-to-RGB) image transformation using CAE. The appraoch could obtain an average SSIM value of 0.81 and RMSE 0.147 for 29,168 images of KAIST dataset. The results obtained from our model for 16,000 patches of KAIST images as shown in Table 4 are marginally better which indicates that our model is giving good performance in comparison to the existing approaches.

To assert that the proposed network can identify similarity between visual and IR image patches, we have performed the following evaluation as done typically in image retrieval kind of applications. The following values are computed:

- A pair of patches is true positive (TP) if it is a corresponding patch pair of VS and IR images and the SSIM value between the encoded representations is greater than a threshold.

- A pair of patches is true negative (TN) if it is not a corresponding patch pair of VS and IR images and the SSIM value between the encoded representations is less than a threshold.
- A pair of patches is false positive (FP) if it is not a corresponding patch pair and the SSIM value is greater than a threshold.
- A pair of patches is false negative (FN) if it is a corresponding patch pair and the SSIM value is less than a threshold.

To measure the quality of encoding, precision, recall, F1 measure and accuracy values are computed using the above mentioned TP, TN, FP, FN values for three different thresholds. In Table 5, we present the results for 0.8, 0.7 and 0.6 thresholds on SSIM similarity score between the patches. The results for all the measures show stable improvement for approach 4a in comparison with approach 1. It is clearly observed that the model is giving precise results in identifying the similar or dissimilar regions of VS and IR images, which is essential for applications such as image matching and retrieval.

5.3 Discussion

The proposed encoder architecture is trained with visual and LWIR patches and tested for patch similarity using two different dataset. From the mentioned results the following inferences can be drawn.

1 The network architecture, hyperparameters are carefully selected based on extensive litearture study of encoder architectures. Activation function is choosen as ELU, BN, and use of SSIM as a loss function, contributed in improving the performance of the network.

2 Reconstruction of a visual patch from IR patch is successfully demonstrated and objective evaluted using SSIM.

3 Multi-modal encoded represenations are most efiicient for image matching when compared to intensity values of patches or unimodal encoded representatives, irrespective of the similairty measure used. This is comprehensively proved using SSIM and EMD.

4 We considered similar and disimiliar patches as two different classes, for which TP, TN, FP, and FN are computed. we obtained an accuracy of 84 perecentage for the propose apparoch. This clearly proves that the encoded representation of mulimodal net are able to distinguish between similar and dissmiliar patches.

5 The proposed approach extends the idea of image trnasformation as proposed in Nyberg et al. (2018) and Berg et al. (2018). Multi-modal CAE can be effectively employed for applications which require similarity computations of Visual and LWIR patches.

6 Conclusions

In this paper, various problems of multi-spectral patch matching are addressed and how deep learning CAE features can give notable results is explored.

We present a 2-channel network architecture, integrated with CAEs for multi-spectral patch matching. One channel is unimodal-net and the other is multi-modal-net. Firstly, our unimodal CAE network is giving promising results in encoding LWIR and visible patches. In order to overcome the spectral differences, we proposed a novel training upon multi-modal-net with CAE as a transformer. This channel is trained with different spectral patches to encode LWIR image patch as VS representative. To obtain a decoded visual patch closest to the original patch, SSIM is used as a loss function to train the network. The model trained with DSSIM is better in comparison with MSE. The encoded representatives are invariant to the spectral differences and similarity between these instance representatives is better in comparison and is able to reduce the differences.

The proposed network architecture is proved using LWIR-visual images and KAIST datasets objectively using MSE, PSNR, EMD and SSIM. The 2-channel transformer-based CAE architecture has greatly improved between LWIR and visual patches as shown in the results. This architecture can further be extended to other multi-modal applications such as retrieval, etc.

Novelty of proposed method is in generating representation of IR patch using corresponding IR and VS patches with an unsupervised CAE architecture. This enables us to employ unimodal similarity measures such as SSIM and EMD for VS-IR patch matching.

In addition to similarity, we have proved the usability of encoded representatives in matching, retrieval.

References

Ali, M.B., Gu, I.Y-H. and Jakola, A.S. (2019) 'Multi-stream convolutional autoencoder and 2D generative adversarial network for glioma classification', in *International Conference on Computer Analysis of Images and Patterns*, pp.234–245.

Appalaraju, S. and Chaoji, V. (2017) *Image Similarity using Deep CNN and Curriculum Learning*, ArXiv abs/1709.08761.

Arif, M. and Mahalanobis, A. (2020) *Multiple View Generation and Classification of Mid-Wave Infrared Images using Deep Learning*, ArXiv abs/2008.07714.

Azarang, A., Manoochehri, H.E. and Kehtarnavaz, N. (2019) 'Convolutional autoencoder-based multispectral image fusion', *IEEE Access*, Vol. 7, pp.35673–35683, DOI: 10.1109/ACCESS.2019.2905511.

Baldi, P. (2012) 'Autoencoders, unsupervised learning, and deep architectures', *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp.37–49.

Balntas, V., Johns, E., Tang, L. and Mikolajczyk, K. (2016) *PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors*, ArXiv abs/1601.05030.

Bank, D., Koenigstein, N. and Giryes, R. (2020) *Autoencoders*, ArXiv abs/2003.05991.

Benbakreti, S., Benouis, M., Roumane, A. and Benbakreti, S. (2021) 'Stacked autoencoder for Arabic handwriting word recognition', *Int. J. Comput. Sci. Eng.*, Vol. 24, No. 6, pp.629–638.

Berg, A., Ahlberg, J. and Felsberg, M. (2018) 'Generating visible spectrum images from thermal infrared', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.1143–1152.

Campo, F.B., Ruiz, F.L. and Sappa, A.D. (2012) 'Multimodal stereo vision system: 3D data extraction and algorithm evaluation', *IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 5, pp.437–446.

Chen, M., Shi, X., Zhang, Y., Wu, D. and Guizani, M. (2017) 'Deep feature learning for medical image analysis with convolutional autoencoder neural network', *IEEE Transactions on Big Data*, Vol. 7, No. 4, pp.750–758.

Cheng, Z., Sun, H., Takeuchi, M. and Katto, J. (2018) 'Deep convolutional autoencoder-based lossy image compression', in *IEEE Picture Coding Symposium (PCS)*, pp.253–257.

Clevert, D-A., Unterthiner, T. and Hochreiter, S. (2015) 'Fast and accurate deep network learning by exponential linear units (ELUS)', *International Conference on Learning Representations (ICLR 2015)*, arXiv preprint arXiv:1511.07289.

Dong, G., Liao, G., Liu, H. and Kuang, G. (2018) 'A review of the autoencoder and its variants: a comparative perspective from target recognition in synthetic-aperture radar images', *IEEE Geoscience and Remote Sensing Magazine*, Vol. 6, No. 3, pp.44–68.

Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L. and Tao, D. (2016) 'Stacked convolutional denoising auto-encoders for feature representation', *IEEE Transactions on Cybernetics*, Vol. 47, No. 4, pp.1017–1027.

Georgiou, T., Liu, Y., Chen, W. and Lew, M. (2020) 'A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision', *International Journal of Multimedia Information Retrieval*, Vol. 9, No. 3, pp.135–170.

Guo, X., Liu, X., Zhu, E. and Yin, J. (2017) 'Deep clustering with convolutional autoencoders', *International Conference on Neural Information Processing*, pp.373–382.

Hani Altwaijry, A.V. and Belongie, S. (2016) 'Learning to detect and match keypoints with deep architectures', in *Proceedings of the British Machine Vision Conference (BMVC)*, pp.49.1–49.12.

Jaques, N., Taylor, S., Sano, A. and Picard, R. (2017) 'Multimodal autoencoder: a deep learning approach to filling in missing sensor data and enabling better mood prediction', in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp.202–208.

Karimpouli, S. and Tahmasebi, P. (2019) 'Segmentation of digital rock images using deep convolutional autoencoder networks', *Computers & Geosciences*, Vol. 126, pp.142–150, ISSN: 0098-3004 [online] http://doi.org/10.1016/j.cageo.2019.02.003.

Kerner, H.R., Wagstaff, K.L., Bue, B.D., Wellington, D.F., Jacob, S., Horton, P., Bell, J.F., Kwan, C. and Amor, H.B. (2020) 'Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions', *Data Mining and Knowledge Discovery*, Vol. 34, No. 6, pp.1642–1675.

Kerner, H.R., Wellington, D.F., Wagstaff, K.L., Bell, J.F., Kwan, C. and Amor, H.B. (2019) 'Novelty detection for multispectral images with application to planetary exploration', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp.9484–9491.

Knyaz, V.A., Vygolov, O., Kniaz, V.V., Vizilter, Y., Gorbatsevich, V., Luhmann, T. and Conen, N. (2017) 'Deep learning of convolutional auto-encoder for image matching and 3D object reconstruction in the infrared range', in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp.2155–2164.

Kolberg, J., Grimmer, M., Gomez-Barrero, M. and Busch, C. (2020) 'Anomaly detection with convolutional autoencoders for fingerprint presentation attack detection', *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 3, No. 2, pp.190–202.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol. 1, ISBN: 9781627480031, DOI: 10.5555/2999134.2999257.

Laakom, F., Raitoharju, J., Iosifidis, A., Nikkanen, J. and Gabbouj, M. (2019) 'Color constancy convolutional autoencoder', in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp.1085–1090.

Li, H. and Wu, X-J. (2018) 'Densefuse: a fusion approach to infrared and visible images', *IEEE Transactions on Image Processing*, Vol. 28, No. 5, pp.2614–2623.

Liu, W., Shen, X., Wang, C., Zhang, Z., Wen, C. and Li, J. (2018) 'H-Net: neural network for cross-domain image patch matching', in *IJCAI*, pp.856–863, DOI: 10.24963/ijcai.2018/119.

Lowe, D.G. (2004) 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision*, Vol. 60, No. 2, pp.91–110.

Luppino, L.T., Hansen, M.A., Kampffmeyer, M., Bianchi, F.M., Moser, G., Jenssen, R. and Anfinsen, S.N. (2020) 'Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images' [online] https://doi.org/10.1109/TGRS.2019.2930348.

Mao, X., Shen, C. and Yang, Y-B. (2016) 'Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections', in *Advances in Neural Information Processing Systems*, Vol. 29, pp.2802–2810 [online] http://arxiv.org/abs/1603.09056.

Masci, J., Meier, U., Cireşan, D. and Schmidhuber, J. (2011) 'Stacked convolutional auto-encoders for hierarchical feature extraction', in *International Conference on Artificial Neural Networks*, pp.52–59.

Meng, Q., Catchpoole, D., Skillicom, D. and Kennedy, P.J. (2017) 'Relational autoencoder for feature extraction', in *2017 International Joint Conference on Neural Networks*, pp.364–371.

Nyberg, A., Bergstrom, D., Petersson, H. and Gustavsson, D. (2018) *Transforming Thermal Images to Visible Spectrum Images using Deep Learning*, p.42 [online] https://www.diva-portal.org/smash/get/diva2:1255342/FULLTEXT01.pdf.

Pawar, K. and Attar, V. (2020) 'Deep learning-based intelligent surveillance model for detection of anomalous activities from videos', *International Journal of Computational Vision and Robotics*, Vol. 10, No. 4, pp.289–311.

Rubner, Y., Tomasi, C. and Guibas, L.J. (2000) 'The Earth mover's distance as a metric for image retrieval', *International Journal of Computer Vision*, Vol. 40, No. 2, pp.99–121.

Shahriari, A. (2016) 'Learning of separable filters by stacked fisher convolutional autoencoders', in *Proceedings of the British Machine Vision Conference (BMVC)*, pp.54.1–54.13.

Simonyan, K. and Zisserman, A. (2014) 'Very deep convolutional networks for large-scale image recognition', *3rd International Conference on Learning Representations, ICLR 2015*.

Song, J., Zhang, H., Li, X., Gao, L., Wang, M. and Hong, R. (2018) 'Self-supervised video hashing with hierarchical binary auto-encoder', *IEEE Transactions on Image Processing*, Vol. 27, No. 7, pp.3210–3221.

Suganuma, M., Ozay, M. and Okatani, T. (2018) 'Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search', in *International Conference on Machine Learning*, pp.4771–4780.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) 'Going deeper with convolutions', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–9.

Tharani, M., Khurshid, N. and Taj, M. (2018) *Unsupervised Deep Features for Remote Sensing Image Matching via Discriminator Network*, ArXiv abs/1810.06470.

Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A. and Tsakalides, P. (2019) 'Survey of deep-learning approaches for remote sensing observation enhancement', *Sensors*, Vol. 19, No. 18, p.3929.

Turchenko, V., Chalmers, E. and Luczak, A. (2017) 'A deep convolutional auto-encoder with pooling-unpooling layers in caffe', *International Journal of Computing*, DOI: 10.48550/ARXIV.1701.04949.

Van Tulder, G. and de Bruijne, M. (2018) 'Learning cross-modality representations from multi-modal images', *IEEE Transactions on Medical Imaging*, Vol. 38, No. 2, pp.638–648.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y. (2014) 'Learning fine-grained image similarity with deep ranking', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Yasrab, R., Gu, N. and Zhang, X. (2017) 'An encoder-decoder based convolution neural network (CNN) for future advanced driver assistance system (ADAS)', *Applied Sciences*, Vol. 7, No. 4, p.312.

Zagoruyko, S. and Komodakis, N. (2015) 'Learning to compare image patches via convolutional neural networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4353–4361.

Zhang, C., Cai, Y., Lin, G. and Shen, C. (2020) 'DeepEMD: few-shot image classification with differentiable Earth mover's distance and structured classifiers', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.12203–12213.

Zhang, Y. (2018) 'A better autoencoder for image: convolutional autoencoder', in *ICONIP17-DCEC* [online] http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf (accessed 23 March 2017).

Zilvan, V., Ramdan, A., Suryawati, E., Kusumo, R.B.S., Krisnandi, D. and Pardede, H.F. (2019) 'Denoising convolutional variational autoencoders-based feature learning for automatic detection of plant diseases', in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pp.1–6.

Žižakić, N., Ito, I., Meeus, L. and Pizurica, A. (2019) 'Autoencoder-learned local image descriptor for image inpainting', in *BNAIC/BENELEARN 2019*, Vol. 2491.