Cross-modal Image Synthesis within Dual-Energy X-ray Security Imagery

Brian K. S. Isaac-Medina[1], Neelanjan Bhowmik[1], Chris G. Willcocks[1], Toby P. Breckon[1,2]
Department of {Computer Science[1], Engineering[2]}, Durham University
Durham, UK
{brian.k.isaac-medina, neelanjan.bhowmik, christopher.g.willcocks, toby.breckon}@durham.ac.uk

Abstract

Dual-energy X-ray scanners are used for aviation security screening given their capability to discriminate materials inside passenger baggage. To facilitate manual operator inspection, a pseudo-colouring is assigned to the effective composition of the material. Recently, paired image to image translation models based on conditional Generative Adversarial Networks (cGAN) have shown to be effective for image colourisation. In this work, we investigate the use of such a model to translate from the raw X-ray energy responses (high, low, effective-Z) to the pseudo-coloured images and vice versa. Specifically, given N X-ray modalities, we train a cGAN conditioned in $N-m$ domains to generate the remaining m representation. Our method achieves a mean squared error (MSE) of 16.5 and a structural similarity index (SSIM) of 0.9815 when using the raw modalities to generate the pseudo-colour representation. Additionally, raw X-ray high energy, low energy and effective-Z projections were generated given the pseudo-colour image with minimum MSE of 2.57, 5.63 and 1.43, and maximum SSIM of 0.9953, 0.9901 and 0.9921. Furthermore, we assess the quality of our synthesised pseudo-colour reconstructions by measuring the performance of two object detection models originally trained on real X-ray pseudo-colour images over our generated pseudo-colour images. Interestingly, our generated pseudo-colour images obtain marginally improved detection performance than the corresponding real X-ray pseudo-colour images, showing that meaningful representations are synthesized and that these reconstructions are applicable for differing aviation security tasks.

1. Introduction

Identification of material composition plays an important role in baggage security screening as it facilitates the material-based detection of prohibited items [16,22]. A material can be characterized by a mass attenuation coefficient which describes how beams at different energy levels are able to penetrate the material. In this sense, multiple energy X-rays can be used to identify the composition of a scanned object. Particularly, dual-energy X-ray imaging has shown to be an effective technique for this task [30]. The effective atomic number, effective-Z or Z_{eff}, can be approximated given two different energy projections between 20 and 200 keV [20]. Subsequently, a look-up table is usually used to assign a material profile and hence corresponding pseudo-colour/RGB to a value of Z_{eff}, identifying between organic (orange), metallic (blue) and inorganic (green) [1]. An example of such X-ray sub-modalities is shown in Fig. 1 where the high and low energy images can be further processed, via the use of effective-Z, to create a corresponding pseudo-coloured images [9].

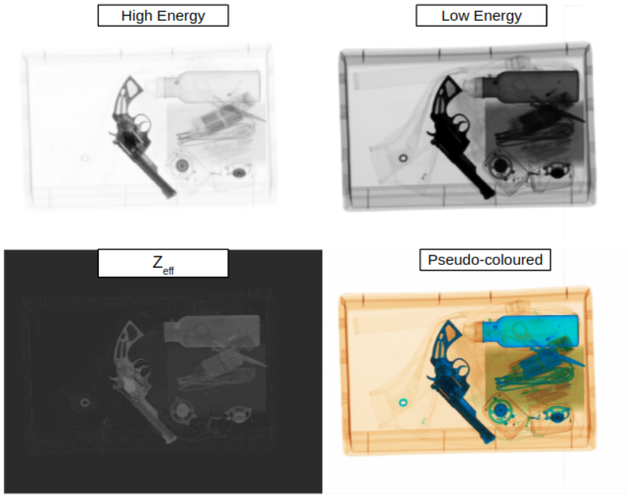


Figure 1. Exemplar multi-modal X-ray screening imagery.

The task of assigning an RGB colour to a greyscale (intensity) value is known as image colourisation. It is an ill-posed problem since the mapping from colour to greyscale $f : \mathbb{R}^3 \to \mathbb{R}$ is not injective; *i.e.* different RGB values may have the same grey value. It has been shown that deep neural networks have good a performance for image colourisation [4]. When paired data is available, a popular supervised architecture for this task is the *pix2pix* architecture proposed by Isola *et al*. [12]. They use a conditional Generative Adversarial Network (cGAN) to generate an image in a different domain than the input image. In this sense, image colourisation is an image to image translation task where the greyscale and the coloured representations of the images are considered to belong to different domains. Since high and low energy responses can be seen as greyscale intensity images, cGAN can be used to translate between energy and coloured images.

Image pseudo-colourisation of dual-energy raw projections has been performed in recent years to aid the visual inspection of security imagery. However, recent works focusing on automatic detection of threat items [2] have brought the question as to whether the raw energy images encode additional information that can be used for this purpose. Bhowmik *et al*. [5] used the raw responses to train different object detection algorithms. They found that the energy responses can be used independently to detect objects of interest, but the best results are obtained when detectors are trained using the pseudo-coloured images, the energy responses and the Z_{eff} mapping in conjunction. Furthermore, they demonstrate that such models are transferable across differing X-ray scanners [5]. Although several large-scale X-ray baggage imagery datasets exist [11, 21, 29], raw X-ray projections are not usually provided as it is not archived by default in standard operational use.

In this context, this work investigates both the generation of pseudo-colour images from dual-energy X-ray security raw modalities (high energy, low energy and Z_{eff}) and the decomposition of these energy images from pseudo-coloured images. Our contributions are as follows:

- use of a GAN-based image to image translation architecture [12] applied to the context of dual-energy X-ray security imagery for the generation of high energy, low energy and Z_{eff} modalities from pseudo-colour X-ray imagery and vice versa.
- the proposed use of two GAN generators for cross-modality synthesis with multiple paired input and output variants, namely, via input concatenation and Siamese network output for each input modality. Maximal quality is obtained with the Siamese version of the generator, with a mean squared error of 16.5 and a structural similarity index measure of 0.9815 for the generation of pseudo-coloured images from the raw X-ray energy modalities.
- assessment of the performance of two object detection models trained on real X-ray imagery when tested on the GAN generated images. Interestingly, the performance on the generated pseudo-colour images outperforms the real X-ray images, showing that meaningful representations are learned with applications in downstream aviation security tasks.

2. Related Work

Earlier image colourisation techniques based on deep learning used plain convolutional neural networks in a supervised fashion [7, 31]. Isola *et al*. [12] proposed the *pix2pix* architecture which uses a cGAN for general paired image-to-image translation tasks. It is demonstrated that cGANs can be used for image colourisation, where the original image and its greyscale version are considered as paired samples. A tailored version of *pix2pix* for image colourisation is explored by Nazeri *et al*. [23]. Image colourisation has also been used to translate from a single-valued domain, such as infrared [18] and radar [25], to a coloured domain. For a comprehensive review on image colourisation, see Anwar *et al*. [4].

Colourisation and enhancement of dual-energy X-ray imagery have been investigated in order to improve detection of threat items [8, 15]. However, to the best of our knowledge, this is the first work that aims to reconstruct the pseudo-colouring image from the raw X-ray projections and to recover the energy responses from the pseudo-colour image.

3. Dual-energy X-ray Imaging

X-ray images are formed by measuring the transmitted irradiance I of a beam with energy E through a material with thickness T and atomic number Z. This resulting irradiance I is given by the Beer's law:

$$I = I_0 e^{-\mu(E,Z)T}\,, \tag{1}$$

where μ is the attenuation coefficient which depends on the material and the energy of the beam. It is noted from Eq. (1) that the transmitted irradiance I is always less or equal to I_0, meaning that thicker objects appear darker in the resulting image, as seen in Fig. 1. Since I_0 and I are known, we can obtain the expression:

$$\mu(E,Z)T = \ln\left(\frac{I}{I_0}\right)\,. \tag{2}$$

For energies less than 200 keV, μ can be decomposed into the attenuation coefficients μ_p and μ_c dominated by the photoelectric and the Compton scattering effects [20], *i.e.*,

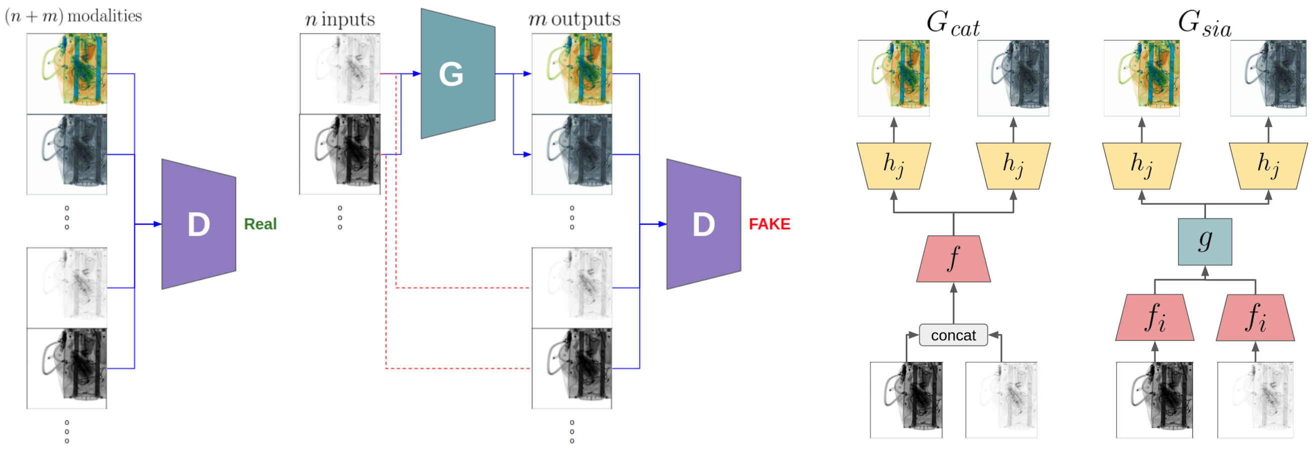
$$\mu(E,Z) = \mu_p(E,Z) + \mu_c(E,Z)\,. \tag{3}$$

Alvarez and Macovski [3] empirically found that:

$$\mu_p(E,Z) \approx \frac{1}{E^3} K_p \frac{\rho}{A} Z^m \tag{4}$$

$$\mu_z(E,Z) \approx f_{KN}(E) K_c \frac{\rho}{A} Z\,, \tag{5}$$

where f_{KN} is the Klein-Nishina function, A is the atomic weight and K_p, K_c and m are constants. An approximation of the atomic number Z can be obtained by using low



(a) (b)

Figure 2. Cross-modal image translation of dual-energy X-ray imagery with a cGAN. (a) Modified *pix2pix* architecture to account for multiple input and outputs. (b) Two generators are proposed: G_{cat} concatenates channel-wise the inputs while G_{sia} has a sub-network for each input modality. Both generators implement different output networks for each output modality.

and high energies E_l and E_h where the response I is dominated by μ_p and μ_c, respectively. Since the response of both energies are measured with respect to the same object, and thus the same thickness, the ratio $\mu_p(E_l, Z)/\mu_c(E_h, Z)$ can be calculated using Eq. (2). From Eqs. (4) and (5), we can express this ratio as:

$$\frac{\mu_p}{\mu_c} \approx \frac{1}{E_l^3 f_{KN}(E_h)} Z^{m-1} . \qquad (6)$$

The atomic number Z is then approximated by:

$$Z \approx K \left(\frac{\mu_p}{\mu_c}\right)^{\frac{1}{n}} , \qquad (7)$$

where K is a value depending on the high and low energies and $n = m - 1$ is a constant. Finally, the thickness of a material can be obtained from Eq. (2).

Since an X-ray beam may penetrate different objects, instead of calculating the Z for each of them, we simplify our analysis by considering that the beam went through an homogeneous material. The resulting atomic number of this hypothetical material is known as the effective atomic number, Z_{eff}. Dual-energy pseudo-coloured images are coloured by assigning a colour depending on the Z_{eff} and the thickness, Eq. (2) [20].

4. Methodology

In this work, we utilise an approach based o the *pix2pix* architecture, modified to account for multiple input and output images, for cross-modality translation of dual-energy X-ray imagery.

4.1. Problem Formulation

The *pix2pix* architecture is a cGAN consisting on a generator $G : \mathbb{R}^{C_{in} \times H \times W} \rightarrow \mathbb{R}^{C_{out} \times H \times W}$ that maps an image with C_{in} input channels and $H \times W$ spatial size from a domain to another domain with the same spatial size and C_{out} output channels, and a discriminator $D : \mathbb{R}^{(C_{in}+C_{out}) \times H \times W} \rightarrow (0, 1)$ that classifies if the image from the target domain is real or fake given the image from the source domain. Given two paired images $\{x_A, x_B\}$ from domains A and B, *pix2pix* uses the adversarial loss function:

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D) = &\mathbb{E}_{x_A, x_B}\left[\log D(x_A, x_B)\right] + \\ &\mathbb{E}_{x_A}\left[\log\left(1 - D(x_A, G(x_A))\right)\right] . \end{aligned} \qquad (8)$$

and additionally, an $\mathcal{L}_{L1}$ reconstruction loss is added as the final image reconstruction objective:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) . \qquad (9)$$

Conventionally, pseudo-coloured X-ray images (e.g. Fig. 1) are formed from the effective atomic number Z_{eff} and the material thickness/density, which are obtained from the high and low energy responses (see Sec. 3). Consequently, we extend the *pix2pix* architecture to accept multiple input and output images in order to allow us to work across the joint set of {*pseudo-colour, high, low,* Z_{eff}} X-ray modalities (as shown in Fig. 1).

4.2. Proposed Variants

Our proposed extended architecture is shown in Fig. 2a. Given n paired inputs $\mathbf{x} = \{x_1, \ldots, x_n\}$ with $\{u_1, \ldots, u_n\}$

channels and m paired outputs $\mathbf{y} = \{y_1, \ldots, y_m\}$ with $\{v_1, \ldots, v_m\}$ channels, we define a multi-domain generator $G : \mathbb{R}^{\sum_i^n u_i \times H \times W} \rightarrow \mathbb{R}^{\sum_i^m v_i \times H \times W}$. Two methods of combining multiple domains are explored in this work: via channel concatenation and via a Siamese network sub-architecture. In the former, the generator G_{cat} takes the input images concatenated channel-wise as a single input for a network f, while in the latter, the generator G_{sia} process each input x_i in a sub-network f_i, where the resulting representations are concatenated channel-wise and combined fed into a common network g. Each domain output y_j is generated from a common feature representation of the input images using a different network h_j for each output modality. A diagram with these approaches is shown in Fig. 2b. The generators G_{cat} and G_{sia} define the generation processes:

$$\begin{aligned} (\mathbf{y}_{\text{cat}})_j &= G_{\text{cat}}(\mathbf{x})_j \\ &= (h_j \circ f)\left([x_1, \ldots, x_n]\right) \end{aligned} \tag{10}$$

and:

$$\begin{aligned} (\mathbf{y}_{\text{sia}})_j &= G_{\text{sia}}(\mathbf{x})_j \\ &= (h_j \circ g)\left([f_1(x_1), \ldots, f_n(x_n)]\right) , \end{aligned} \tag{11}$$

where $[\ldots]$ means concatenation. Similarly to the discriminator in the *pix2pix* architecture, our multi-domain discriminator $D : \mathbb{R}^{(\sum_i u_i + \sum_j v_j) \times H \times W} \rightarrow (0, 1)$ takes all inputs and outputs to classify them as real or fake.

The multi-domain adversarial and reconstruction losses are then:

$$\begin{aligned} \mathcal{L}_{mcGAN}(G, D) = & \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\log D(\mathbf{x}, \mathbf{y})\right] + \\ & \mathbb{E}_{\mathbf{x}}\left[\log\left(1 - D(\mathbf{x}, G(\mathbf{x}))\right)\right] \end{aligned} \tag{12}$$

and:

$$\mathcal{L}_{mL1}(G) = \sum_{j}^{m} \mathbb{E}_{\mathbf{x},y_j}\left[\|y_j - G(\mathbf{x})_j\|_1\right] . \tag{13}$$

Furthermore, Jiang *et al.* [13] introduced the frequency focal loss (FFL), which aims to reduce the gap in the frequency response of the synthesized images. We investigate if our multi-domain image translation can be improved using the FFL. Finally, our objective function is then:

$$\begin{aligned} G^* = \arg\min_G \max_D & \mathcal{L}_{mcGAN}(G, D) + \\ & \lambda_{mL1}\mathcal{L}_{mL1}(G) + \lambda_{\text{FFL}}\mathcal{L}_{\text{FFL}}(G) . \end{aligned} \tag{14}$$

4.3. Network Architecture

The original *pix2pix* model uses a UNet [24] with skip connections as the generator. However, following the approach of CycleGAN [32], we implement the architecture described by Johnson *et al.* [14]. This network consists on three convolutional layers, a series of stacked residual blocks, two transposed convolutional layers and an output convolutional layer. Following this architecture, Tab. 1 describes the f_i, g and h_j networks used for the generators in Eqs. (10) and (11). The f_i networks consist on three convolutional layers and L residual blocks, the g network is composed of M residual blocks and the h_j networks have N residual blocks, two transposed convolutions and a final convolutional layer. All layers use instance normalisation [26] and ReLU activation except for the last convolutional layer in h_j, that does not use normalisation and has a Tanh function as activation. The network f in Eq. (10) is defined as $f = g \circ f_i$. In this work we have three different cases of cross-modal synthesis: one-to-one mode, multi-to-one mode and one-to-multi modes. For one-to-one and one-to-many task we use $L = 4$, $M = 5$ and $N = 0$ while for many-to-one we use $L = M = N = 3$. Finally, the discriminator follows the PatchGAN network used by Isola *et al.* [12].

Network	Architecture
f_i	Conv 7×7 Conv 3×3, stride = 2 Conv 3×3, stride = 2 $L\times$ Residual
g	$M\times$ Residual
h_j	$N\times$ Residual Transp Conv 3×3, stride = 2 Transp Conv 3×3, stride = 2 Conv, 7×7

Table 1. Architecture of the generator sub-networks.

5. Evaluation

We evaluate our multi-modal cross-modal translation architecture for pseudo-coloured and raw X-ray energy response images (as shown in Fig. 1). We use the labels *rgb*, *h*, *l* and *z* for the pseudo-colour, high energy, low energy and Z_{eff} imagery subsets, respectively. The experiments performed in this work are described in Tab. 2.

5.1. Dataset

We train our models in the *deei6* dataset [5]. This dataset consists on 7,022 quadruplets (h, l, z and rgb) of bags scanned in a dual-energy Gilardoni FEP ME 640 AMX scanner [10] (see Fig. 1). Bounding box and instance mask annotations are given for six classes: bottle, hairdryer, iron, toaster, phone-tablet and laptop. The dataset is split in 4,909 quadruplets for training and 2,113 for testing.

Experimental label	Reconstruction Type	Description
$\{h, l, z\} \rightarrow rgb$	one-to-one	High energy, low energy or Z_{eff} to pseudo colour.
$hl_{\text{sia}} \rightarrow rgb$	many-to-one	High and low energy to pseudo colour (G_{sia}).
$hlz_{\text{sia}} \rightarrow rgb$	many-to-one	High energy, low energy and Z_{eff} to pseudo colour (G_{sia}).
$hlz_{\text{cat}} \rightarrow rgb$	many-to-one	High energy, low energy and Z_{eff} to pseudo colour (G_{cat}).
$rgb \rightarrow \{h, l, z\}$	one-to-one	Pseudo colour to high energy, low energy or Z_{eff}
$rgb \rightarrow hlz$	one-to-many	Pseudo colour to high energy, low energy and Z_{eff}

Table 2. Experimental labels and descriptions for the experiments carried out in this work.

Model	w/o FFL		w/ FFL	
	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑
$h \rightarrow rgb$	182.3	0.9229	185.6	0.9216
$l \rightarrow rgb$	183.6	0.9296	168.4	0.9297
$z \rightarrow rgb$	125.0	0.9041	121.5	0.9049
$hl_{\text{sia}} \rightarrow rgb$	465.1	0.9411	101.6	0.9600
$hlz_{\text{cat}} \rightarrow rgb$	20.1	0.9753	18.9	0.9766
$hlz_{\text{sia}} \rightarrow rgb$	**16.5**	**0.9815**	17.3	0.9808
$rgb \rightarrow h$	2.57	**0.9953**	**2.28**	0.9948
$rgb \rightarrow \boldsymbol{h}lz$	18.3	0.9910	12.3	0.9915
$rgb \rightarrow l$	5.63	**0.9901**	**5.43**	0.9888
$rgb \rightarrow h\boldsymbol{l}z$	7.55	0.9847	6.13	0.9885
$rgb \rightarrow z$	38.0	0.9823	4.61	0.9830
$rgb \rightarrow hl\boldsymbol{z}$	1.43	0.9794	**0.76**	**0.9921**

Table 3. Cross-modality reconstruction performance.

5.2. Performance Metrics

Two image quality metrics are used in this work: mean squared error (MSE) and the structural similarity index measure (SSIM) [28]. Additionally, two detection networks, CARAFE [27] and Cascade Mask RCNN [6], are trained on the real X-ray image datasets using the same settings as Bhowmik *et al*. [5] and tested on the synthesized images generated from the same X-ray dataset under the experimental conditions set out in Tab. 2. We report instance segmentation results using the MS COCO mean Average Precision (mAP) performance metric [19] (intersection over union of 0.50:.05:0.95), using Average Precision (AP) for class-wise and mAP for overall performance.

5.3. Implementation Details

Input images are resized to 600 × 600 pixels and random cropped to have a final size of 512 × 512. Differently from *pix2pix*, we do not use dropout. The model is trained using Adam optimization [17] with a learning rate of 2×10^{-4} for 100 epochs, linearly decaying to 0 for another 100 epochs. We choose $\lambda_{L1} = 100$ for the objective function defined in Eq. (14) and $\lambda_{\text{FFL}} = 10$ when the FFL is used. A batch size of 6 n-tuples of image modalities is used to train our models.

6. Results

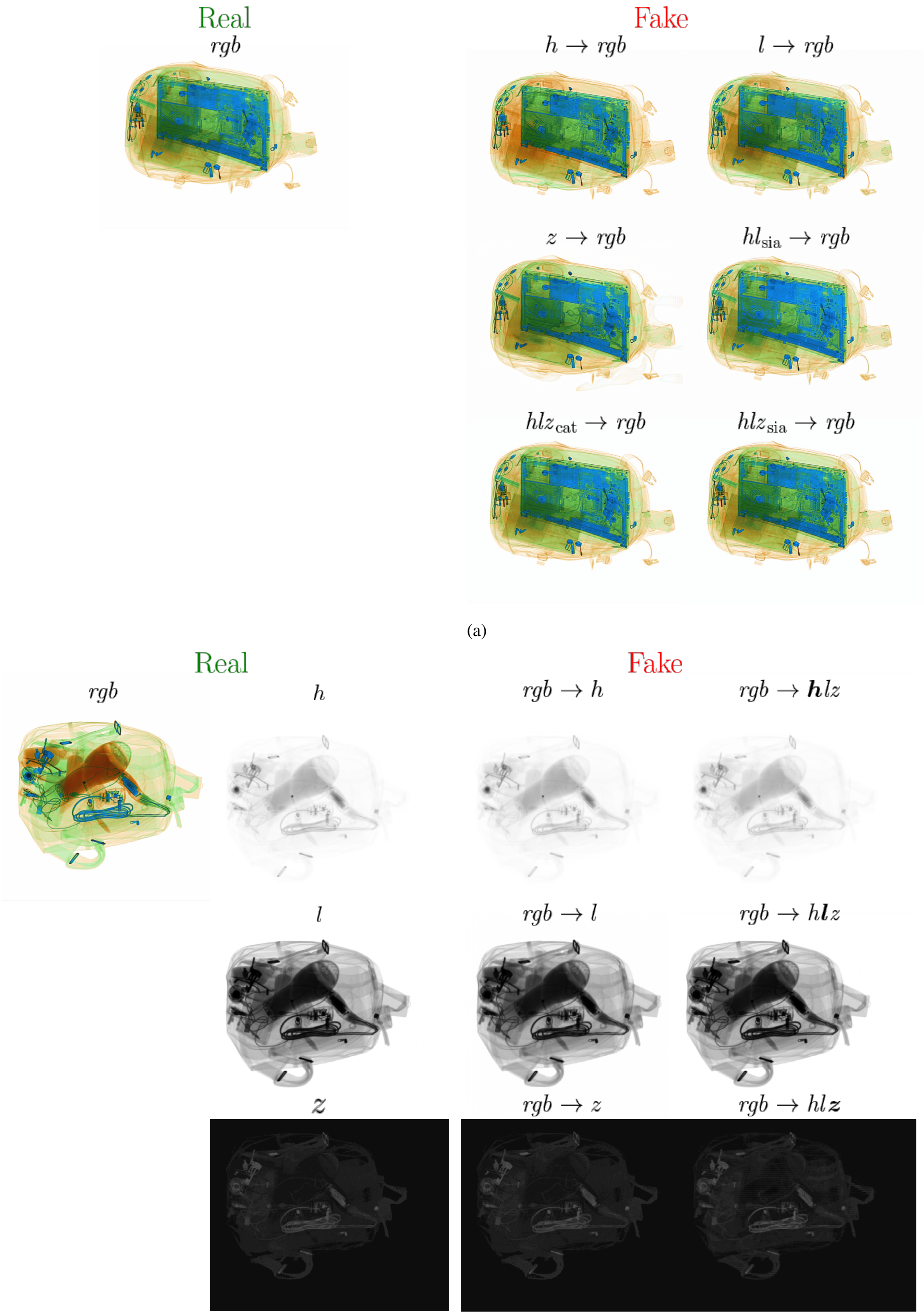
In this section we review the results for image synthesis quality and detection performance. We evaluate the G_{sia} and G_{cat} generators and the impact of the FFL during training.

6.1. Reconstruction Quality

Cross-modality image synthesis performance is shown in Tab. 3. MSE and SSIM metrics are reported, comparing the synthesis quality with the real images. The impact of using the FFL is also reported. It can be observed that in general, the best reconstructions are obtained when using the focal frequency loss, although the improvement is minor and does not always lead to the best results.

The best pseudo-coloured reconstructions are obtained by using the three modalities h, l and z and the G_{sia} generator from Eq. (11), obtaining an MSE of 16.5 and SSIM of 0.9815. We also confirm that pseudo-coloured image reconstruction gets degraded when only using one energy level. Although the use of Z_{eff} individually significantly improves the MSE, the structural similarity gets worse because the thickness information is lost (Sec. 3). Fig. 3a shows an example of the pseudo-colour reconstructions. It can be seen that when only using the high or low energy images, the reconstructed image tends to get confused around the organic (orange) regions, getting materials mixed up. Although the material information can be matched better using only the Z_{eff} modality, the shape is not always obtained correctly (see for example the top right corner of the laptop). It is also observed that using more than just one modality creates very accurate reconstructions.

As seen in Tab. 3, the energy modalities can be recovered with high SSIM from the pseudo-colour images. The best results for high and low energies are obtained when they are generated using separate models. This could be explained by earlier layers learning specific features that capture the effect from each energy modality. However, the Z_{eff} modality is better recovered when predicting the three raw modalities at the same time, meaning that the learned features guided from the other modalities help in the identification of the atomic number. Fig. 3b shows an example of the high energy, low energy and Z_{eff} modalities synthesized


Real
rgb
Fake
$h \rightarrow rgb$
$l \rightarrow rgb$
$z \rightarrow rgb$
$hl_{\text{sia}} \rightarrow rgb$
$hlz_{\text{cat}} \rightarrow rgb$
$hlz_{\text{sia}} \rightarrow rgb$
Real
rgb
h
Fake
$rgb \rightarrow h$
$rgb \rightarrow \boldsymbol{h}lz$
l
$rgb \rightarrow l$
$rgb \rightarrow h\boldsymbol{l}z$
$\boldsymbol{z}$
$rgb \rightarrow z$
$rgb \rightarrow hl\boldsymbol{z}$

(a)

(b)

Figure 3. Exemplar of cross-modality synthesis. (a) Raw modalities to pseudo-colour. (b) pseudo-colour to raw modalities.

Dataset		Bottle	Hairdryer	Iron	Toaster	P-tablet	Laptop	mAP
High Energy	Real	**0.641/0.628**	**0.640/0.657**	**0.675/0.689**	**0.787/0.793**	**0.516/0.533**	**0.771/0.776**	**0.672/0.679**
	$rgb \rightarrow h$	0.597/0.593	0.579/0.594	0.642/0.656	0.740/0.760	0.496/0.498	0.754/0.751	0.635/0.642
	$rgb \rightarrow h$ (FFL)	0.596/0.591	0.584/0.596	0.632/0.655	0.738/0.756	0.469/0.481	0.741/0.747	0.627/0.638
	$rgb \rightarrow \boldsymbol{h}lz$	0.578/0.571	0.548/0.552	0.613/0.638	0.728/0.745	0.476/0.469	0.733/0.744	0.613/0.620
	$rgb \rightarrow \boldsymbol{h}lz$ (FFL)	0.590/0.584	0.553/0.563	0.618/0.641	0.715/0.724	0.480/0.474	0.737/0.747	0.615/0.622
Low Energy	Real	**0.615/0.620**	**0.609/0.629**	**0.657/0.682**	0.751/**0.776**	**0.508/0.526**	**0.760/0.765**	**0.650/0.666**
	$rgb \rightarrow l$	0.585/0.606	0.552/0.569	0.626/0.649	0.747/0.762	0.498/0.490	0.731/0.743	0.623/0.637
	$rgb \rightarrow l$ (FFL)	0.584/0.607	0.559/0.574	0.630/0.651	**0.759**/0.771	0.507/0.500	0.739/0.748	0.630/0.642
	$rgb \rightarrow h\boldsymbol{l}z$	0.559/0.563	0.524/0.545	0.605/0.637	0.740/0.751	0.471/0.462	0.706/0.716	0.601/0.612
	$rgb \rightarrow h\boldsymbol{l}z$ (FFL)	0.578/0.593	0.544/0.561	0.623/0.649	0.740/0.758	0.494/0.493	0.727/0.733	0.618/0.631
Z_{FFL}	Real	0.534/**0.548**	**0.460/0.490**	**0.606**/0.634	**0.783/0.793**	**0.490/0.488**	**0.718**/0.732	**0.598/0.614**
	$rgb \rightarrow z$	0.533/0.540	0.355/0.386	0.604/**0.635**	0.779/0.786	0.485/0.483	**0.718**/0.732	0.579/0.593
	$rgb \rightarrow z$ (FFL)	**0.535**/0.543	0.442/0.471	0.603/0.634	0.776/0.787	0.483/0.480	0.715/**0.736**	0.592/0.609
	$rgb \rightarrow hl\boldsymbol{z}$	0.472/0.494	0.290/0.304	0.544/0.560	0.745/0.756	0.403/0.395	0.642/0.666	0.516/0.529
	$rgb \rightarrow hl\boldsymbol{z}$ (FFL)	0.460/0.492	0.241/0.271	0.551/0.576	0.766/0.767	0.387/0.391	0.611/0.616	0.502/0.519
Pseudo Colour	Real	0.638/**0.635**	0.609/0.638	0.662/0.694	0.788/0.790	**0.536/0.552**	0.754/0.776	0.665/0.681
	$h \rightarrow rgb$	0.575/0.573	0.517/0.528	0.557/0.576	0.718/0.729	0.419/0.441	0.703/0.722	0.581/0.595
	$h \rightarrow rgb$ (FFL)	0.567/0.567	0.512/0.532	0.557/0.573	0.715/0.730	0.424/0.445	0.716/0.730	0.582/0.596
	$l \rightarrow rgb$	0.525/0.534	0.290/0.315	0.423/0.432	0.704/0.719	0.388/0.398	0.520/0.577	0.475/0.496
	$l \rightarrow rgb$ (FFL)	0.556/0.569	0.396/0.400	0.503/0.494	0.734/0.747	0.430/0.435	0.615/0.671	0.539/0.553
	$z \rightarrow rgb$	0.560/0.554	0.476/0.478	0.571/0.577	0.777/0.784	0.480/0.482	0.748/0.756	0.602/0.605
	$z \rightarrow rgb$ (FFL)	0.568/0.566	0.489/0.487	0.572/0.578	0.779/0.790	0.484/0.482	0.743/0.754	0.606/0.609
	$hl_{\mathrm{sia}} \rightarrow rgb$	0.513/0.514	0.454/0.456	0.583/0.539	0.726/0.732	0.420/0.425	0.478/0.476	0.529/0.524
	$hl_{\mathrm{sia}} \rightarrow rgb$ (FFL)	0.615/0.615	0.531/0.531	0.660/0.642	0.783/0.791	0.479/0.485	0.727/0.738	0.632/0.634
	$hlz_{\mathrm{cat}} \rightarrow rgb$	0.634/0.627	0.628/0.639	0.678/0.697	0.792/0.799	0.517/0.532	0.771/0.773	0.670/0.678
	$hlz_{\mathrm{cat}} \rightarrow rgb$ (FFL)	0.635/0.628	0.621/0.636	0.683/0.700	0.792/0.795	0.531/0.544	0.769/0.772	0.672/0.679
	$hlz_{\mathrm{sia}} \rightarrow rgb$	0.637/0.631	**0.637/0.649**	**0.688/0.704**	**0.793/0.802**	0.524/0.537	**0.773/0.777**	**0.675/0.683**
	$hlz_{\mathrm{sia}} \rightarrow rgb$ (FFL)	**0.641/0.635**	0.628/0.644	0.685/0.701	**0.793/0.802**	0.528/0.536	0.768/**0.777**	0.674/0.682

Table 4. Object detection results using different modalities of X-ray imagery from the *deei6* dataset. The two reported values are for the CARAFE [27] and Cascade Mask RCNN [6] architectures.

from the pseudo-colour image. Some small blurring effects can be seen in the high and low energy generations for the $rgb \rightarrow hlz$ model. Nevertheless, it is seen that regardless the model, the generated images exhibit high fidelity.

6.2. Detection Performance

Detection performance for real and synthesized images is presented in Tab. 4. Results are for instance segmentation predictions. They are presented with two values, each corresponding to the CARAFE and Cascade Mask RCNN models. Per-class AP and total mAP results are shown.

Synthesized raw modalities show a better detection performance when they are generated with individual models, which is consistent with the quality of the reconstructions in Tab. 3. Compared to the real images, the detection performance in the synthesized raw modalities gets reduced. This means that although the generated images may seem very similar, the reconstructions do not perfectly match the energy projections. It is worth noticing that while the generated Z_{eff} from the $rgb \rightarrow hlz$ shows a good SSIM, its detection performance is reduced significantly while compared to the original Z_{eff} response. This shows that detection models are very sensitive to small variations in the input images. On the other hand, the mAP of the generated pseudo-colour rgb images gets improved by a 1% for CARAFE detection model when using the three raw modalities and the G_{sia} generator. This slight improvement over the detection performance may indicate that our model is learning to generate pseudo-coloured images more effectively than the standard formulation in terms of information retention in the resulting pseudo-coloured visualisation. These results illustrate that our proposed approach can be used to learn meaning from representations across differing X-ray modalities such that they can be used to effectively train a secondary deep neural network for subsequent downstream tasks.

7. Conclusions

In this work we investigate the use of a conditional generative adversarial network for image to image translation of dual-energy X-ray security imagery. We perform image colourisation from high energy, low energy and effective atomic number Z_{eff} modalities and vice versa. Two novel generator architectures are proposed for the combination of multiple modalities as inputs and outputs. The first generator, G_{sia}, takes each input into a sub-network and then concatenates the resulting features. Our second proposed generator, G_{cat}, concatenates channel-wise the input images and process it as a single image multi-channel input. In both cases, multiple outputs are generated by having a subnetwork to generate each modality. The use of the focal frequency loss (FFL) is also investigated.

It is observed that the best results for image colourisation are obtained when using the three modalities (high energy, low energy and Z_{eff}) and the G_{cat} generator, achieving a SSIM of 0.9766. In general, the FFL improved image colourisation. The best results for the extraction of the high and low energy modalities are obtained when having a separate model for each, having SSIMs of 0.9953 and 0.9901 (without FFL). On the other hand, the Z_{eff} gets a better reconstruction when using a model that predicts the three raw modalities at the same time, achieving a SSIM of 0.9921. A qualitative assessment shows that the differences are barely noticeable and reconstruction exhibit a good similarity when compared to the original X-ray modality imagery.

Detection performance results were obtained for two different architectures trained on the real images and tested on the synthesized images. For the raw X-ray energy response imagery, performance is worse on the generated images and compared to the original imagery. However, the pseudo-coloured images generated using the three raw modalities and the G_{sia} generator show a better detection performance than that obtained for the real images. On this basis, we hypothesize that the model learnt for raw X-ray energy response to pseudo-colour image translation offers a superior mapping in terms of information retention than the original raw X-ray imagery.

Future work will investigate the use of modern architectures for higher definition image to image translation and the transferability of these model to images obtained from different scanners that have no raw X-ray energy data availability.

Acknowledgments

This work is partially supported by the Mexican Council of Science and Technology (CONACyT).

References

[1] Besma Abidi, Y. Zheng, Andrei Gribok, and Mongi Abidi. Screener evaluation of pseudo-colored single energy x-ray luggage images. In *Proc. of Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 35–35, 2005. 1

[2] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022. 2

[3] R E Alvarez and A Macovski. Energy-selective reconstructions in x-ray computerised tomography. *Physics in Medicine and Biology*, 21(5):733–744, sep 1976. 2

[4] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *arXiv preprint arXiv:2008.10774*, 2020. 1, 2

[5] N. Bhowmik, Y.F.A. Gaus, and T.P. Breckon. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In *Proc. Int. Conf. on Image Processing*, pages 1224–1228. IEEE, September 2021. 2, 4, 5

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *CoRR*, abs/1906.09756, 2019. 5, 7

[7] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proc. of the IEEE International Conference on Computer Vision*, pages 415–423, 2015. 2

[8] Mohamed Chouai, Mostafa Merah, José-Luis Sancho-Gomez, and Malika Mimi. Dual-energy x-ray images enhancement based on a discrete wavelet transform fusion technique for luggage inspection at airport. In *International Conference on Image and Signal Processing and their Applications*, pages 1–6, 2019. 2

[9] Krzysztof Dmitruk, Michal Mazur, Marcin Denkowski, and Pawel Mikolajczak. Method for filling and sharpening false colour layers of dual energy x-ray images. *IFAC-PapersOnLine*, 48(4):342–347, 2015. 13th IFAC and IEEE Conference on Programmable Devices and Embedded Systems. 1

[10] Gilardoni. Fep me 640 amx, 2018. "https://www.gilardoni.it/en/security/x-ray-solutions/automatic-detection-of-explosives/fep-me-640-amx/". (accessed: 22.03.2022). 4

[11] Lewis D Griffin, Matthew Caldwell, and Jerone T A Andrews. Compass-xp. In *Zenodo*, May 2019. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proc. of Computer Society Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4

[13] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for generative models. *CoRR*, abs/2012.12821, 2020. 4

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016. 4

[15] R. Kayalvizhi, Amit kumar, S. Malarvizhi, Anita Topkar, and P. Vijayakumar. Raw data processing techniques for material classification of objects in dual energy x-ray baggage inspection systems. *Radiation Physics and Chemistry*, 193:109512, 2022. 2

[16] Sajid Ullaha Khan, Imran Ullahb Khan, Imdadc Ullah, Naveedd Saif, and Irfane Ullah. A review of airport dual energy x-ray baggage inspection techniques: Image enhancement and noise reduction. *Journal of X-Ray Science and Technology*, 28:481–505, 06 2020. 1

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015. 5

[18] Matthias Limmer and Hendrik PA Lensch. Infrared colorization using deep convolutional neural networks. In *IEEE International Conference on Machine Learning and Applications*, pages 61–68. IEEE, 2016. 2

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5

[20] Harry E Martz and Steven M Glenn. Dual-energy x-ray radiography and computed tomography. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2019. 1, 2, 3

[21] Caijing Miao, Lingxi Xie, Fang Wan, chi Su, Hongye Liu, jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proc. of the IEEE International Conference on Computer Vision*, 2019. 2

[22] A. Mouton and T.P. Breckon. A review of automated image understanding within 3d baggage computed tomography security screening. *Journal of X-Ray Science and Technology*, 23(5):531–555, September 2015. 1

[23] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International conference on articulated motion and deformable objects*, pages 85–94. Springer, 2018. 2

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 4

[25] Qian Song, Feng Xu, and Ya-Qiu Jin. Radar image colorization: Converting single-polarization to fully polarimetric using deep neural networks. *IEEE Access*, 6:1647–1661, 2018. 2

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 4

[27] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. CARAFE: content-aware reassembly of features. *CoRR*, abs/1905.02188, 2019. 5, 7

[28] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[29] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and deocclusion attention module. In *Proc. of the International Conference on Multimedia*, pages 138–146, 2020. 2

[30] Zhengrong Ying, Ram Naidu, and Carl Crawford. Dual energy computed tomography for explosive detection. *Journal of X-Ray Science and Technology*, 14:235–256, 01 2006. 1

[31] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 4