Quad-DIP for X-ray cargo image decomposition

Zheng Hu Qiang Li Gang Fu
AI Center, Nuctech Company Limited
huzheng@nuctech.com liqiang1@nuctech.com fugang@nuctech.com

Yuxiang Xing Li Zhang
Tsinghua University
xingyx@mail.tsinghua.edu.cn zli@mail.tsinghua.edu.cn

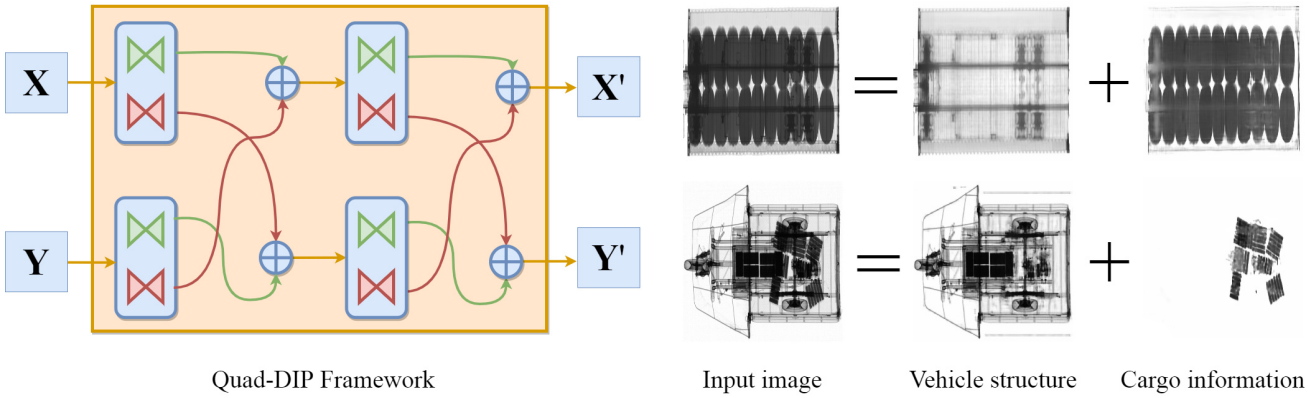

Figure 1: Quad-DIP. Left: The Quad-DIP network trained with two cargo vehicle images of same vehicle type as inputs and their reconstructions as outputs. Right: X-ray images of cargo vehicles decomposed into two parts: vehicle structure part and cargo information part. The first row is example in container cargo inspection scenario, the second row is another scenario for tricycle application.

Abstract

To identify different cargoes on vehicles accurately in scanned image is a tough issue. An unsupervised image decomposition method, based on a novel dual-stage double-DIP (DDIP) network, named as Quad-DIP, was proposed for the decomposition of X-ray scanned image of a cargo vehicle into vehicle and goods separately without ground truth data. The model could be effectively trained based on the fact that, firstly, the structure contents of same type vehicles were similar in the images, and secondly, the contents of goods on different vehicles were different and independent to each other. Our work focus on the content-wise correlation between them. The vehicle structure could be identified from two inputs containing the same type of vehicles, and the image could be decomposed into two components of vehicle structure and cargo information accurately after the training of Quad-DIP. We examine the accuracy of this method on the collected X-ray cargo vehicle dataset. The decomposition of Quad-DIP was more accurate than those of other published methods in literature.

1. Introduction

Image decomposition, to decompose one image to two different images, is one of the important tasks in computer vision. Components derived may be different at different situation. For example, one can represent the structure information and the other the texture information. In the image de-noising task, one can be considered as signal and the other as noise. As for image segmentation, one may represent foreground, and the other background [6, 2, 21]. In this study, X-ray images of vehicles loaded with all kinds of goods are required to be decomposed into two parts: vehicle structure and cargo information, to facilitate the task of cargo categorization (Fig. 1) during automatic X-ray image

inspection [18].

There are two challenges in this work. Firstly, no empty vehicle images can be obtained to serve as supervision labels for vehicle structure learning, so only an unsupervised approach can be employed in our work. Secondly, the detail of vehicle components is so complex that it may vary from image to image even they are of a same type. Besides the none strict positioning of vehicles between an x-ray source and detector, there are many deformable parts such as tires, tank and engine, vehicle maintenance and modification, as well as other dynamic factors uncontrollable, it is difficult to learn an exact model to represent generic structure information in vehicle images. Although many studies in literature so far shed light on dealing with this kind of problems, the application scenarios of the current methods are not suitable for this task due to the inconsistency of background in cargo vehicle images.

As an image decomposition task, the parts relating to the structure of same type vehicles are similar, or can be predicted to some extent, while the parts relating to cargo are different and hard to be foreseen. We employ this strategy to constrain the decomposition conditions in this study, namely, to train a group of images containing same type vehicles, minimizing the distance of common parts and maximizing the distance between different parts. However, it is difficult for the model to learn the complex and diverse individual differences of the vehicle structure due to that there is no accurate supervised information of the vehicle structure corresponding to each image, which greatly affects the decomposition effect. We solve this problem by constructing a dual-stage decomposition network.

The key of the proposed solution lies in three aspects as shown in Fig.1. Firstly, a dual-stage cross-generation network is designed. The influence of in-class differences of vehicle structures can be reduced by using two decomposition stages and two cross-mixing stages with two inputs of images. In this process, we refer to the idea of CycleGAN [26], where a two-tier architecture is designed to make the data generated cyclically from the domain $\mathcal{D}_1$ to domain $\mathcal{D}_2$ and then from domain $\mathcal{D}_2$ back to domain $\mathcal{D}_1$, and the RealNVP model [3], where a cross recombination structure to connect the dual-stage networks is designed. Secondly, we use a combined loss of mean-square error (MSE) and the perceptual loss function [10] to measure the accuracy of reconstructed images, and use the local cross-correlation (LCC) function to measure the structural similarity of the decomposed components [1]. Finally, the input images are aligned by Sift [16] and Demons [22] registration.

The proposed method is applied to a practical dataset of real X-ray cargo vehicle images that has 3897 images of two vehicle types (Fig. 2).

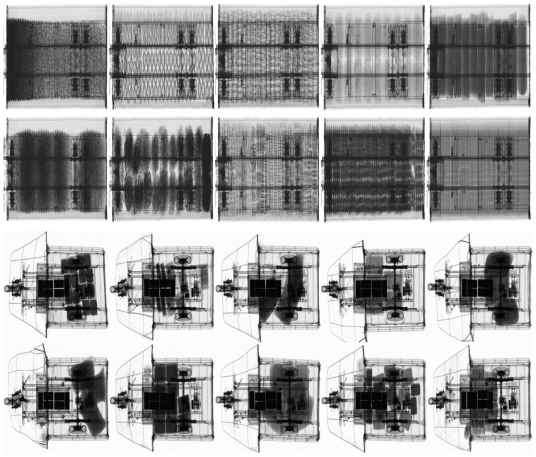

Figure 2: X-ray vehicle dataset.

2. Related work

Unsupervised image decomposition is a challenging task. Many related methods have been published in the literature. We briefly review these works according to different application scenarios and implementation forms.

From the perspective of application scenarios. The input data of the network is normally limited by different application scenarios and thus different algorithms have been proposed. For decomposition network based on single image input scenario, "DDIP", proposed by Gandelsman *et al.* [6] is pioneering research. The input could be decomposed to two-parts through a DDIP structure which has two deep_image_prior (DIP) blocks [21]. The detailed content of the decomposed part could be retained by reconstruction loss, and the independence of each component is limited by correlation loss. Rendered Intrinsics Network (RIN) [9] showed that it could make use of reconstruction loss to improve its intermediate representations by learning both the image decomposition and recombination functions, which allows one to use unlabeled data during training. Subr *et al.* [20] developed an algorithm for decomposing images into multiple scales of superposed oscillations based on the key observation that the spatial scale of oscillations could be characterized by the density of local extrema. After studying the statistics of natural images in the Labelme dataset, Gai *et al.* [5] not only confirmed the well-known sparsity of image gradients, but also discovered new joint behavior patterns of image gradients. Based on these statistical properties, they developed a sparse blind separation algorithm to estimate both layer motions and linear mixing coefficients and finally recover all layers. For the decomposition network based on several images or video sequence, the inputs were obtained under a single scene. The image sequence has rather static background, the foreground and background content can be learned much accurate and ef-

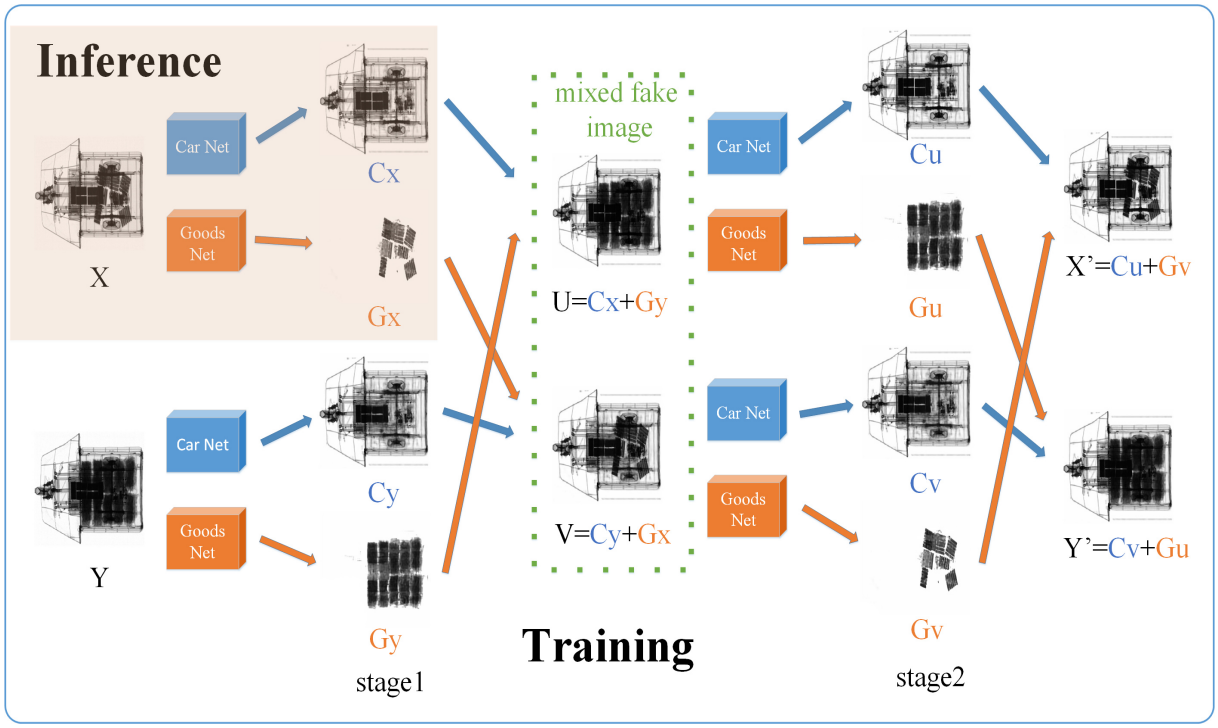

Figure 3: Quad-DIP network architecture.

fectively [13, 24, 7, 11, 12, 23]. Faktor *et al.* [4] combines the task of image segmentation with decomposition by extracting the commonness of a set of images. Lin *et al.* [14] proposed an effective image decomposition method for anomaly detection (AD) based on dual deep reconstruction networks (DDR-ID). The proposed method aims to achieve normal-class-specific image decomposition in an end-to-end manner by optimizing for AD oriented objectives together with image reconstruction.

From network design aspect. Generate models such as VAE and GAN have gained tremendous attentions and have been applied in a wide range. In the field of image decomposition, many methods were designed with this idea. Li and Snavely [13] proposed an image decomposition method based on AE, by which the image was decomposed into two layers through one encoding and two decoding processes, constrained by minimizing the distance between the first layer and maximizing the distance between the corresponding second layer. DADNet [27] used a generator to generate decomposed component (2 output parts), and use a discriminator to judge whether the result image is a clean image or a mixed image. Liu *et al.* [15] proposed a method combining VAE and GAN, based on the knowledge that natural images and their reflected and shaded images had the same unchanging content. It was believed that the image mixing with reflected light and shadow just had a change in the image domain. Based on this idea, the authors converted the image domain with the same content instead of estimating the reflection/shadow parts from a naturally mixed image. The domain features were trained by collected reflection/shadow images.

Ma *et al.* [17] established a cross-recombination architecture to accomplish an intrinsic image decomposition task. The model had a single stage architecture, and the reconstructed image can be well restored because of the consistency of illumination. However, it is complex and hard to handle in this task because even the same type vehicles could have significant differences from each other due to the factors of geometrical inconsistency in scanning and many deformable parts. The "middle images" formed after a single reorganization (such as **U**, **V** in Fig. 1) cannot be simply reconstructed with the original images **X** and **Y**. The ambiguity of vehicle structure will exist based on such a single stage architecture. Further decomposition and reorganization to generate $\mathbf{X}'$ and $\mathbf{Y}'$ can help to solve this problem. Hence, we designed a dual-stage reconstruction architecture in order to gain a good performance on the decomposition of the vehicle structure in X-ray images. There is still no such design in literature so far.

3. Method

In this work, we are to achieve decompose the X-ray image of a cargo vehicle into two components, namely the vehicle structure and the cargo information. Based on the

fundamental physical law of X-ray photon attenuation, the attenuation of each ray can be modeled as the summation of attenuation from materials along the ray path. Hence, an X-ray image $\mathbf{X}$ of a vehicle can be formulated as

$$\mathbf{X} = \mathbf{C} + \mathbf{G} \tag{1}$$

with $\mathbf{C}$ denoting the attenuation from vehicle itself and $\mathbf{G}$ denoting the attenuation from good on the vehicle.

Normally empty vehicles (vehicle without any load) are not scanned so that images of empty vehicles are not available to us. Due to the lack of ground truth, the model cannot be trained in a supervised learning manner or a discriminator manner like GAN. To deal this problem, we construct a method to use multiple images (loading different kind of goods) of same type vehicles to learn vehicle features to learn vehicle features. For example, if image $\mathbf{X}$ and image $\mathbf{Y}$ contain the same type vehicle but different cargo, the decomposed part $\mathbf{C}$ from $\mathbf{X}$ and $\mathbf{Y}$ should be similar but different in part $\mathbf{G}$. Seen as Eq. (2),

$$\mathbf{X} = \mathbf{C}_x + \mathbf{G}_x \tag{2-1}$$

$$\mathbf{Y} = \mathbf{C}_y + \mathbf{G}_y \tag{2-2}$$

$$\mathbf{C}_x = \mathbf{C}_y \tag{2-3}$$

$$\mathbf{G}_x \neq \mathbf{G}_y \tag{2-4}$$

The Eq. (2) can be solved by methods published in [13, 24, 23, 17]. However, in the reality of our tasks $\mathbf{C}_x$ and $\mathbf{C}_y$ could be quite different, namely,

$$\mathbf{C}_x \approx \mathbf{C}_y \tag{3}$$

In our work, we try to solve this problem by introducing ambiguity, meanwhile eliminating it using a new dual-stage DDIP network structure named Quad-DIP. On the basis of DDIP structure, the two decomposed components can be achieved through the dual-stage intersecting generation network structure. The typical procedure is described in the following four sections. The dual-stage network architecture designed in Quad-DIP is described in Section 3.1, the Loss function in Section 3.2, the optimization formula derivation in Section 3.3, and the implementation details including image preprocessing and alignment in Section 3.4.

3.1. Quad-DIP Architecture

As shown in Fig. 3, in the training phase, two input images ($\mathbf{X}$ and $\mathbf{Y}$) from one type of vehicles are fed into the network. There are two stages in this architecture, and each stage consists of two DDIP blocks and one cross-mixing block. The DDIP block decomposes an image into corresponding sub-components, while the cross-mixing recombines the sub-components from different images to generate images of mixed information. In the inference phase, we

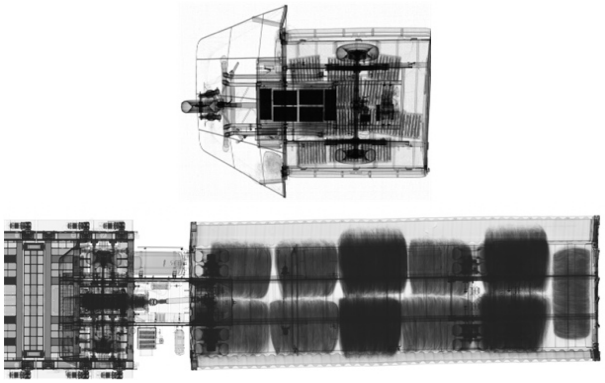

Figure 4: Two kinds of vehicles. Top: cargo tricycle image. Bottom: large container truck image.

can obtain the decomposition result through a single image. Since the four DDIPs share parameters with each other, we can use the first DDIP decomposition result in stage one.

Dual-stage DDIP block: In the proposed Quad-DIP network, there are four DDIP blocks in total. Each DDIP includes one DIP structure for vehicle image extraction (the CarNet) and another DIP for cargo image extraction (GoodsNet), as follows,

$$\begin{aligned} DDIP(\mathbf{X}) &= DIP_{CarNet}(\mathbf{X}) + DIP_{GoodsNet}(\mathbf{X}) \\ &= \mathbf{C}_x + \mathbf{G}_x \end{aligned} \tag{4}$$

The parameters of CarNet and GoodsNet are not shared. The input images $\mathbf{X}$ and $\mathbf{Y}$ as well as the mixed images $\mathbf{U}$ and $\mathbf{V}$ are decomposed to corresponding subcomponents, respectively. The process can be represented by Eq. (5).

$$DDIP(\mathbf{X}) = \mathbf{C}_x + \mathbf{G}_x \tag{5-1}$$

$$DDIP(\mathbf{Y}) = \mathbf{C}_y + \mathbf{G}_y \tag{5-2}$$

$$DDIP(\mathbf{U}) = \mathbf{C}_u + \mathbf{G}_u \tag{5-3}$$

$$DDIP(\mathbf{V}) = \mathbf{C}_v + \mathbf{G}_v \tag{5-4}$$

The four CarNets and four GoodsNets share parameters, respectively. A good DDIP decomposition needs to meet two conditions: Firstly, the result of vehicle structure component (component $\mathbf{C}$) should be as similar as possible, and the result of cargo information component (component $\mathbf{G}$) should be as different as possible. Secondly, the original image can be reconstructed by the addition of component $\mathbf{C}$ and component $\mathbf{G}$.

Cross-mixing block: The input of stage1 is the original images $\mathbf{X}$ and $\mathbf{Y}$, and the output is the mixed images $\mathbf{U}$ and $\mathbf{V}$. $\mathbf{U}$ and $\mathbf{V}$ are mixed fake image generated by cross-recombining the vehicle structure information and cargo information from the original images $\mathbf{X}$ and $\mathbf{Y}$ in stage1, respectively, through the cross-mixing block. See Eq. (6),

$$\mathbf{U} = \mathbf{C}_x + \mathbf{G}_y, \quad \mathbf{V} = \mathbf{C}_y + \mathbf{G}_x \tag{6}$$

where $\mathbf{U}$ is obtained by linear addition of $\mathbf{C}_x$ (from image $\mathbf{X}$) and $\mathbf{G}_y$ (from image $\mathbf{Y}$), and $\mathbf{V}$ is obtained by linear addition of $\mathbf{C}_y$ (from image $\mathbf{Y}$) and $\mathbf{G}_x$ (from image $\mathbf{X}$).

Stage 2 takes the output of stage 1 as its input, that is, the mixed image $\mathbf{U}$ and $\mathbf{V}$ obtained in Eq. 6. Further decomposing images $\mathbf{U}$ and $\mathbf{V}$ by the DDIP in the stage 2 gives rise to sub-components $\mathbf{C}_u$, $\mathbf{G}_u$, $\mathbf{C}_v$ and $\mathbf{G}_v$ of the mixed fake image (Eqs. 5-3, 5-4). Then the reconstructed image $\mathbf{X}'$ and $\mathbf{Y}'$ can be obtained by further cross-mixing, see Eq. (7).

$$\mathbf{X}' = \mathbf{C}_u + \mathbf{G}_v, \quad \mathbf{Y}' = \mathbf{C}_v + \mathbf{G}_u \tag{7}$$

In the next sub-section, we introduce the constraints for the network training and the design of loss functions in detail.

3.2. Loss Function

Quad-DIP network trains a good decomposition model through the following constraints: **Constraint 1**, within the single DDIP, the decomposed parts $\mathbf{C}$ and $\mathbf{G}$ are different from each other; **Constraint 2**, addition of the two components decomposed by DDIP can reconstruct the original image; **Constraint 3**, part $\mathbf{C}$ (the vehicle structure part) from different DDIPs should be close; **Constraint 4**, part $\mathbf{G}$ (the cargo information part) from different vehicle images should be as different as possible; **Constraint 5**, since the mixed fake image is derived from the decomposition results of $\mathbf{X}$ and $\mathbf{Y}$, the decomposition results of $\mathbf{C}$ and $\mathbf{G}$ in different stages should be consistent with each other; **Constraint 6**, the $\mathbf{X}'$ and $\mathbf{Y}'$ generated after the final reorganization of the two stages should be the same as the original input images $\mathbf{X}$ and $\mathbf{Y}$.

Based on the above six principles, the loss function can be expressed as Eq. (8),

$$\begin{aligned} Loss = Loss_{Recon} + \alpha \times Loss_{car\_car} + \beta \times \\ Loss_{car\_goods} + \gamma \times Loss_{goods\_goods} \end{aligned} \tag{8}$$

where the first item is the reconstruction loss, the other three terms are similarity losses with α, β, γ being hyper parameters ranging from 0 to 1. In this study, we just use MSE for the reconstruction loss and LCC function to calculate the similarity losses.

Reconstruction loss: There are eight reconstruction losses in this item, which are calculated by the L2-loss in this study. The formula can be written as follows, $Loss_{Recon}(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|$, where $\mathbf{A}$ and $\mathbf{B}$ are two arbitrary images with same size. In addition, the perceptual loss function is also added to the reconstruction losses because it is considered to be more accurate to measure the similarity between images than the per-pixel loss function [10, 25]. Instead of calculating on the final image, multiple feature maps are adopted to calculate it according to the idea of perceptual loss.

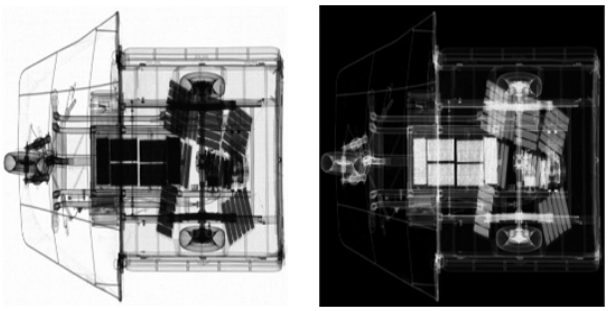

Figure 5: log and reverse processing of X-ray image. Left: original image. Right: processed image.

Based on the principle of constraint 2, four losses are defined as $Loss_{Recon}(\mathbf{X}, \mathbf{C}_x + \mathbf{G}_x)$, $Loss_{Recon}(\mathbf{Y}, \mathbf{C}_y + \mathbf{G}_y)$, $Loss_{Recon}(\mathbf{X}, \mathbf{X}')$ and $Loss_{Recon}(\mathbf{Y}, \mathbf{Y}')$, and based on the constraint 5, the other four losses are defined as: $Loss_{Recon}(\mathbf{C}_x, \mathbf{C}_u)$, $Loss_{Recon}(\mathbf{C}_y, \mathbf{C}_v)$, $Loss_{Recon}(\mathbf{G}_x, \mathbf{G}_v)$ and $Loss_{Recon}(\mathbf{G}_y, \mathbf{G}_u)$

Similarity loss: The similarity loss includes three losses which are the similarity between cars ($Loss_{car\_car}$), the difference between car and goods ($Loss_{car\_goods}$), and the difference between goods ($Loss_{goods\_goods}$).

As we shown in Fig. 3, there are four DDIP structures, so we have four $Loss_{car\_car}$, which are calculated by the LCC loss function [1]. The value of LCC function ranges from 0 to 1, and bigger value indicates more similar images . So we can formula the $Loss_{car\_car}$ as follow, $Loss_{car\_car}(\mathbf{A}, \mathbf{B}) = 1 - LCC(\mathbf{A}, \mathbf{B})$. Based on the principle of constraint 3, the four losses are defined as $Loss_{car\_car}(\mathbf{C}_x, \mathbf{C}_y)$, $Loss_{car\_car}(\mathbf{C}_u, \mathbf{C}_v)$, $Loss_{car\_car}(\mathbf{C}_x, \mathbf{C}_v)$ and $Loss_{car\_car}(\mathbf{C}_y, \mathbf{C}_u)$.

There are four $Loss_{car\_goods}$, which are calculated by the loss function, $Loss_{car\_goods}(\mathbf{A}, \mathbf{B}) = LCC(\mathbf{A}, \mathbf{B})$ to calculate it. Based on the principle of constraint 1, the four losses are defined as $Loss_{car\_goods}(\mathbf{C}_x, \mathbf{G}_x)$, $Loss_{car\_goods}(\mathbf{C}_y, \mathbf{G}_y)$, $Loss_{car\_goods}(\mathbf{C}_u, \mathbf{G}_u)$ and $Loss_{car\_goods}(\mathbf{C}_v, \mathbf{G}_v)$

There are four $Loss_{goods\_goods}$, which are calculated by the loss function, $Loss_{goods\_goods}(\mathbf{A}, \mathbf{B}) = LCC(\mathbf{A}, \mathbf{B})$ to calculate it. Based on the principle of constraint 4, the four losses are defined as $Loss_{goods\_goods}(\mathbf{G}_x, \mathbf{G}_y)$, $Loss_{goods\_goods}(\mathbf{G}_u, \mathbf{G}_v)$, $Loss_{goods\_goods}(\mathbf{G}_x, \mathbf{G}_u)$ and $Loss_{goods\_goods}(\mathbf{G}_y, \mathbf{G}_v)$.

3.3. Feasibility derivation

To train a DDIP structure to decompose the image $\mathbf{X}$ and $\mathbf{Y}$ into two parts, respectively, can be represented by Eq. (9-1) and (9-2),

$$\begin{aligned} DDIP(\mathbf{X}) = \mathbf{C}_x + \mathbf{G}_x, \\ \mathbf{C}_x = \mathbf{GT}_{cx} + \mathbf{E}_x, \ \mathbf{G}_x = \mathbf{GT}_{gx} - \mathbf{E}_x \end{aligned} \tag{9-1}$$

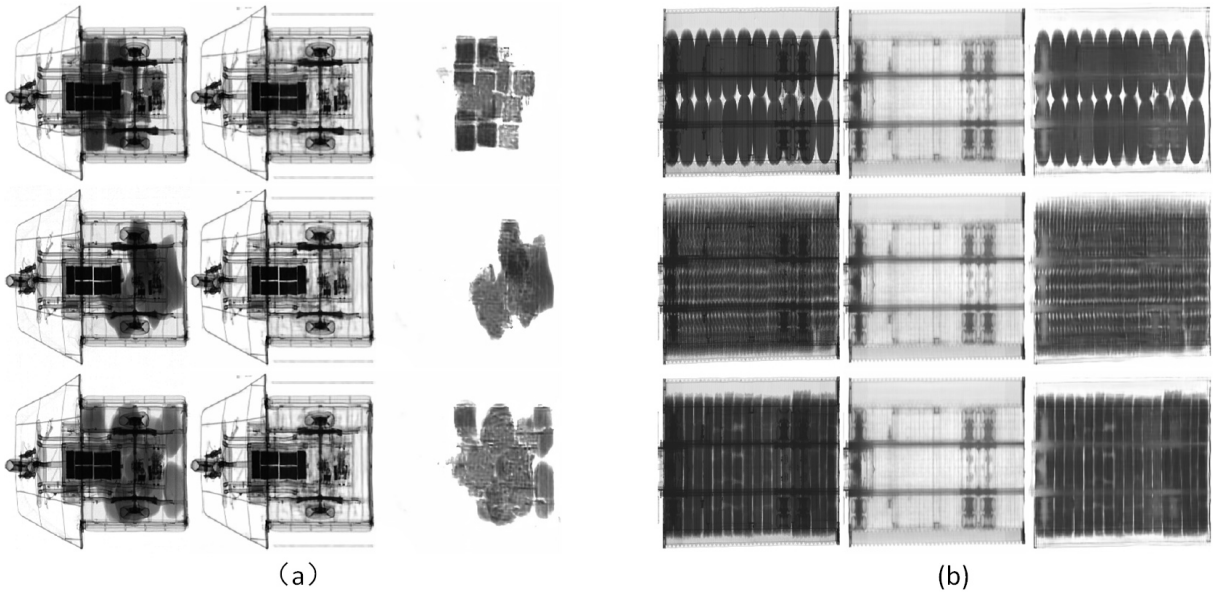

Figure 6: Typical decomposition results from two type of vehicles: tricycle (a) and large container truck (b). In each row of the results (a) and (b) are, from left to right, the original image ($\mathbf{X}$), the vehicle structure ($\mathbf{C}_x$), and the cargo information ($\mathbf{G}_x$)

$$\begin{aligned} DDIP(\mathbf{Y}) &= \mathbf{C}_y + \mathbf{G}_y, \\ \mathbf{C}_y = \mathbf{GT}_{cy} + \mathbf{E}_y,\ \mathbf{G}_y &= \mathbf{GT}_{gy} - \mathbf{E}_y \end{aligned} \tag{9-2}$$

where $\mathbf{GT}_{cx}$ and $\mathbf{GT}_{gx}$ are defined as the ground truth of the vehicle part and that of the goods part of image $\mathbf{X}$, respectively, and $\mathbf{E}_x$ is the content of the decomposition error or ambiguity to some extent (the content belonging to vehicle is mis-divided to the part of goods and vis versa). In the field of image decomposition, the optimization goal of any algorithm is to make $\mathbf{E}_x$ go to zero. The Quad-DIP network constructs a dual-stage structure to achieve this goal.

First of all, the decomposition formula of original image $\mathbf{X}$, $\mathbf{Y}$, mixed fake image $\mathbf{U}$, $\mathbf{V}$ and reconstructed image $\mathbf{X}'$, $\mathbf{Y}'$ can be written by Eq. (10).

$$DDIP(\mathbf{X}) = (\mathbf{GT}_{cx} + \mathbf{E}_x) + (\mathbf{GT}_{gx} - \mathbf{E}_x) \tag{10-1}$$

$$DDIP(\mathbf{Y}) = (\mathbf{GT}_{cy} + \mathbf{E}_y) + (\mathbf{GT}_{gy} - \mathbf{E}_y) \tag{10-2}$$

$$DDIP(\mathbf{U}) = (\mathbf{GT}_{cu} + \mathbf{E}_u) + (\mathbf{GT}_{gu} - \mathbf{E}_u) \tag{10-3}$$

$$DDIP(\mathbf{V}) = (\mathbf{GT}_{cv} + \mathbf{E}_v) + (\mathbf{GT}_{gv} - \mathbf{E}_v) \tag{10-4}$$

In constraint 5 what described in section3.2, the vehicle structure of $\mathbf{U}$ is partly from image $\mathbf{X}$, so $\mathbf{C}$ parts of image $\mathbf{X}$ and $\mathbf{U}$ should be equal, namely the first term on the right hand side of the Eqs. (10-1) and (10-3) should be equal. Given that $\mathbf{GT}_{cx} = \mathbf{GT}_{cu}$, this constraint is equivalent to Eq. (11).

$$\mathbf{E}_x = \mathbf{E}_u \tag{11}$$

Likewise, combine Eqs. (10-2) and (10-4), we obtain Eq. (12).

$$\mathbf{E}_y = \mathbf{E}_v \tag{12}$$

Combine Eqs. (6), (9-1) and (9-2), we obtain Eq. (13).

$$\mathbf{U} = \mathbf{C}_x + \mathbf{G}_y = (\mathbf{GT}_{cx} + \mathbf{E}_x) + (\mathbf{GT}_{gy} - \mathbf{E}_y) \tag{13}$$

In constraint 2 described in section3.2, the sum of the decomposed two parts should be equal to the original, $\mathbf{U} = DDIP(\mathbf{U})$. Combining Eqs. (10-3) and (13), we get Eq. (14).

$$\begin{aligned} (\mathbf{GT}_{cx} + \mathbf{E}_x) + (\mathbf{GT}_{gy} - \mathbf{E}_y) = (\mathbf{GT}_{cu} + \mathbf{E}_u) \\ + (\mathbf{GT}_{gu} - \mathbf{E}_u) \end{aligned} \tag{14}$$

The vehicle structure of $\mathbf{U}$ is partly from image $\mathbf{X}$, and the cargo information is partly from image $\mathbf{Y}$. So $\mathbf{GT}_{cu} = \mathbf{GT}_{cx}$, $\mathbf{GT}_{gu} = \mathbf{GT}_{gy}$. From Eq. (14), we obtain

$$\mathbf{E}_x = \mathbf{E}_y \tag{15}$$

Therefore, combine Eqs. (11), (12) and (15), we obtain Eq. (16).

$$\mathbf{E}_x = \mathbf{E}_u = \mathbf{E}_v = \mathbf{E}_y \tag{16}$$

That is, $\mathbf{E}_x$, $\mathbf{E}_y$, $\mathbf{E}_u$, $\mathbf{E}_v$ are the direct current (DC) component in given dataset, and should be a part of ground truth in vehicle structure component. In the training procedure, they should be approached to zero along with loss descending.

4. Experiments

4.1. Setup

Datasets. This method is tested on a vehicle X-ray image dataset with 3897 images, including 2769 cargo tricycle images and 1128 large container truck images. All of them are collected from a top view, as shown in Fig.4.

Image process. Compared with visual image, X-ray image has its unique properties. According to Beer's law of X-ray attenuation, the collect signals from X-ray scan systems need pre-processed to fit for the training of deep-learning network.

a. Negative logarithm processing. X-ray image is the image produced by the transmitted signals after passing through scanned objects. Different structures of an object are revealed due to the different attenuation property of various substances. The grayscale of an X-ray image, that we refer as transparency in physics, is defined as the ratio of the intensity of the transmitted intensity of X-rays after passing through the scanned object to the incident intensity from X-ray sources. It can be expressed as

$$T = \frac{I}{I_0} = e^{-\int u(l)dl} \tag{17}$$

where $u(l)$ is the linear mass attenuation coefficient, and l is the coordinate on a ray path with dl denoting the small mass thickness of the material on the ray path. Due to the exponential function, the accumulated linear attenuation of different materials along a ray path, e.g., with two overlapped objects can be formulated as $u(l)$, with being the linear attenuation from objects a and b. Hence, we apply a negative logarithm operation when calculating reconstruction loss of the decomposed image:

$$-log(T_{a+b}) = -log(T_a) - log(T_b) \tag{18}$$

An example result of the negative logarithm process is shown in Fig. 5 for illustration.

b. Image alignment. To ease training, we apply an image alignment step in our data pre-processing since there is no strict control of the positioning of vehicles during X-ray scan of cargo vehicles that leads to differently transformed image. In the traditional CV methods, sift is often used for rigid matching and then the dense registration is used for non-rigid matching. In addition, the Spatial Transformer Networks (STN) is also widely used for adaptive registration in recent years [8]. In our work, we use sift and dense to align the data before training manually.

Implementation details: We use the U-net in our generate network [19]. The Loss function that we introduced in Eq. (8) has three parameters α, β, γ, representing the weight of $Loss_{car\_car}$ similarity, $Loss_{car\_goods}$ difference and $Loss_{goods\_goods}$ difference, respectively. In our experiments, we set α=2.5e-4, β=3.5e-5, γ=3.5e-5. The input image is uniformly scaled to 256×256 pixels. We implement the network on Pytorch. The network was trained 50K epochs with Adam optimizer on a Nvidia's Tesla V100 graphics card. It took about 24 hours.

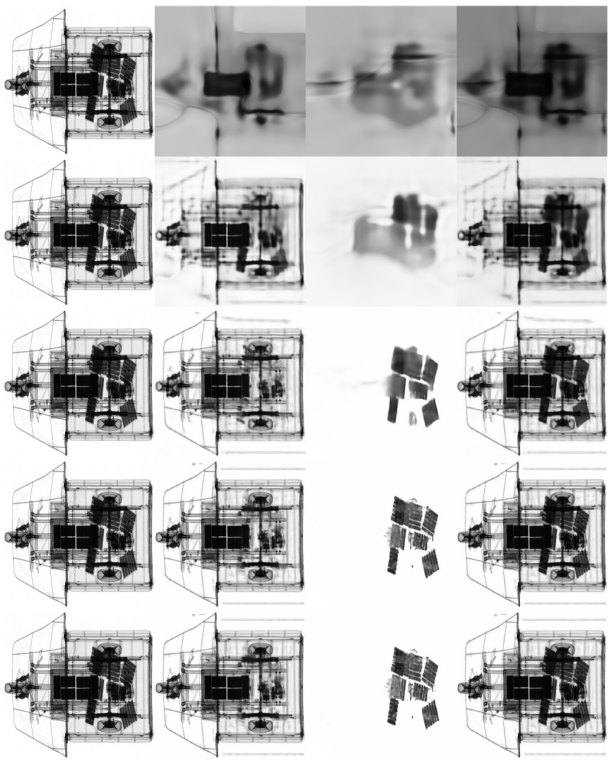

Figure 7: The decomposition result in different epoch during the training process.

Our works solve the problem to separate similar structure part and different cargo part from images of various systems at different sites. Because of the uniqueness of this application scenario, we have not found any other suitable data set for the time being, so we have not tested the effect of our algorithm on any open-access datasets. In the future work, we are considering extending the method to X-ray medical images, and then we might be able to test it on more datasets.

4.2. Results

In this section, we validate the performance and demonstrate the decomposition effects of our method.

Fig. 6 shows the decomposition result of two typical kinds of cargo vehicle images. The left are the results of cargo tricycles, and the right are the results of container truck. In each row, the original images, the vehicle structure component, and the cargo information component are shown from left to right. It can be seen that the decomposition algorithm can separate vehicle structure and the goods well, with details of goods information well preserved. We

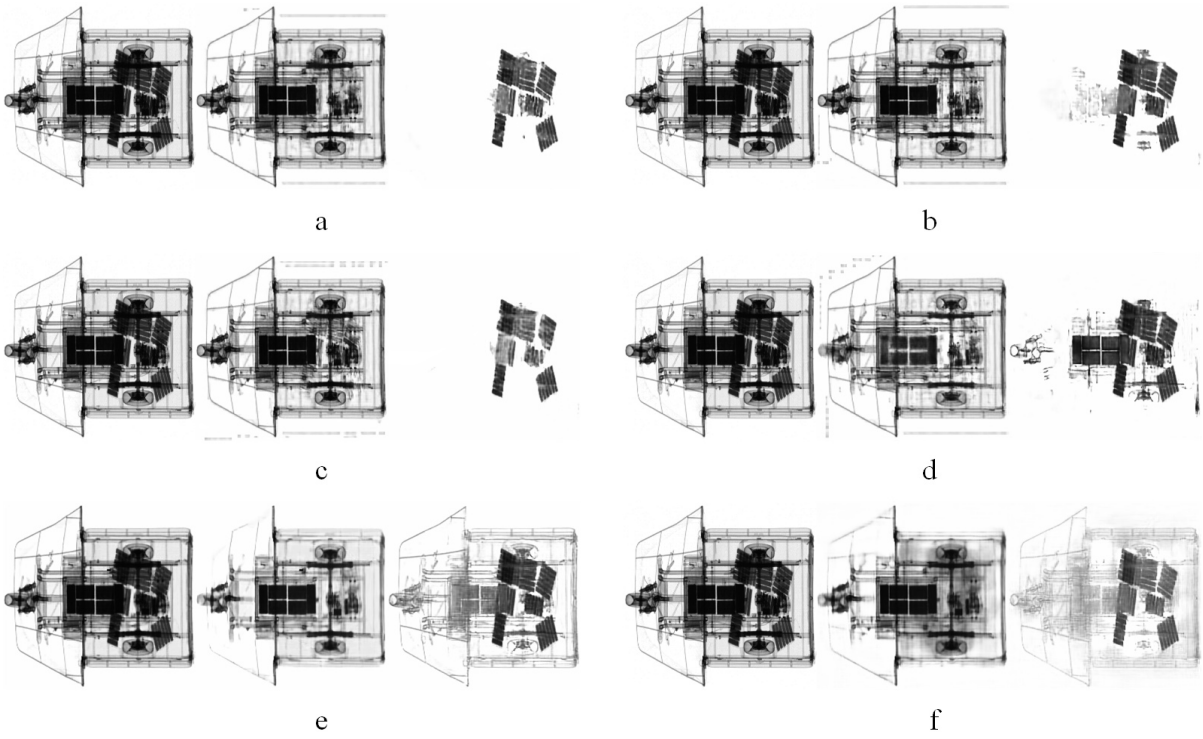

Figure 8: Results from ablation studies and comparison of results from different methods. In each group of results, from left to right are the original image ($\mathbf{X}$), the vehicle structure part ($\mathbf{C}_x$) and the cargo information part ($\mathbf{G}_x$). (a): ours. (b): ours with no image alignment. (c): ours with L1 loss for reconstruction and LCC loss for similarity. (d): ours with L2 loss for reconstruction and orthogonal loss for similarity. (e): DDIP. (f): Wei-Chiu Ma's method.

also find that the algorithm has a relatively poor performance in the region where X-ray is of low penetration (the area with very low pixel value). This is expectable because very little information is in this region, neither vehicle structure or cargo information is clear.

Fig. 7 shows five intermediate results during the training. From top to bottom, they are decomposition results at epoch 100, 1000, 5000, 20000 and 50000 respectively. In each row, from left to right are the original image ($\mathbf{X}$), the vehicle structure part ($\mathbf{C}_x$), the cargo information part ($\mathbf{G}_x$) and the reconstruction image ($\mathbf{X}'$).

A representative decomposition result with proposed Quad-DIP method is shown in Fig. 8a. In each group of results, from left to right are the original image ($\mathbf{X}$), the vehicle structure part ($\mathbf{C}_x$) and the cargo information part ($\mathbf{G}_x$). To further examine the proposed method, we did some ablation studies: 1) use images with no preprocessing of alignment as inputs; 2) use L1 loss for reconstruction instead of L2; 3) use orthogonal loss instead of LCC loss for similarity. The results from Quad-DIP with these three variations are shown in Fig. 8b, 8c, and 8d. As we can see that skipping alignment step will leads to some residual cross-talk in decomposed components. We can see similar phenomena with orthogonal loss for similarity metric. L1 reconstruction loss gives very similar results as L2 reconstruction loss. We also implemented DDIP and Wei-Chiu Ma's method [17] for comparison. By comparing their results in Fig. 8e and 8f with Fig. 8a, we can see that Quad-DIP shows the best effect in our task.

5. Conclusion

We propose a method to decompose an X-ray image of a cargo vehicle into a vehicle structure part and a cargo information part with a deep learning network that can be trained unsupervisely. The model uses a dual-stage decomposition architecture to get accurate image components. In the future work, for more complex decomposition tasks, we can try deeper network structure. Moreover, if with some images of known lables, GAN discriminator can be added to further constrain the quality of decomposition results.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *MI*, 2019. 2, 5

[2] S. Bagon O. Boiman and M. Irani. What is a good image segment? a unified approach to segment extraction. *ECCV*, 2008. 1

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *Machine Learning*, 2017. 2

[4] Alon Faktor and Michal Irani. Co-segmentation by composition. *ICCV*, 2013. 3

[5] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *PAMI*, 2011. 2

[6] Yossi Gandelsman, Assaf Shocher, and Michal Irani. Double-dip": Unsupervised image decomposition via coupled deep-image-priors. *CVPR*, 2018. 1, 2

[7] Daniel Hauagge, Scott Wehrwein, Kavita Bala, and Noah Snavely. Photometric ambient occlusion. *CVPR*, 2013. 3

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CVPR*, 2015. 7

[9] Michael Janner, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, and Joshua B. Tenenbaum. Self-supervised intrinsic image decomposition. *CVPR*, 2018. 2

[10] Johnson Justin, Alahi Alexandre, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. 2, 5

[11] Pierre-Yves Laffont and Jean-Charles Bazin. Intrinsic decomposition of image sequences from local temporal variations. *ICCV*, 2015. 3

[12] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics*, 2012. 3

[13] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. *CVPR*, 2018. 3, 4

[14] Dongyun Lin, Yiqun Li, Shudong Xie, Tin Lay Nwe, and Sheng Dong. Ddr-id: Dual deep reconstruction networks based image decomposition for anomaly detection. *CVPR*, 2020. 3

[15] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. *CVPR*, 2020. 3

[16] David G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999. 2

[17] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. *ECCV*, 2018. 3, 4, 8

[18] Domingo Mery. X-ray testing by computer vision. *CVPRW*, 2013. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CVPR*, 2015. 7

[20] Kartic Subr, Cyril Soler, and Fr´edo Durand. Edge-preserving multiscale image decomposition based on local extrema. *ACM Transactions on Graphics*, 2009. 2

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *CVPR*, 2018. 1, 2

[22] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. *MICCAI*, 2007. 2

[23] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Learning of structure and motion from video. *CVPR*, 2017. 3, 4

[24] Yair Weiss. Deriving intrinsic images from image sequences. *ICCV*, 2001. 3, 4

[25] Xuaner Zhang, Ng Ren, and Qifeng Chen. Single image reflection separation with perceptual losses. *CVPR*, 2018. 5

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 2

[27] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. *CVPR*, 2020. 3