UNSUPERVISED HARMONIOUS IMAGE COMPOSITION FOR DISASTER VICTIM DETECTION

N. Zhang[1,*], F. Nex[1], G. Vosselman[1], N. Kerle[1]

[1] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands
(n.zhang, f.nex, george.vosselman, n.kerle)@utwente.nl

Commission III, WG III/IVa

KEY WORDS: Deep Learning, Victim Detection, Composite Image Generation, Unsupervised Deep Harmonization, Disaster Management

ABSTRACT:

Deep detection networks trained with a large amount of annotated data achieve high accuracy in detecting various objects, such as pedestrians, cars, lanes, *etc*. These models have been deployed and used in many scenarios. A disaster victim detector is very useful when searching for victims who are partially buried by debris caused by earthquake or building collapse. However, considering that larger quantities of real images with buried victims are difficult to obtain for training, a deep detector model cannot give full play to its advantages. In this paper we generate realistic images for training a victim detector. We first randomly cut out human body parts from an open source human data set and paste them into the ruins background images. Then, we propose an unsupervised generative adversarial network (GAN) to harmonize the body parts to fit the style (illumination, texture and color characteristics) of the background. These generated images are finally used to fine-tune a detection network YOLOv5. We evaluate both the AP (average precision) for IoU (Intersection over Union) 0.5 and for IoU $\in$ [0.5:0.05:0.95], which are denoted as $AP@0.5$ and $AP@[.5 : .95]$, respectively. The best experimental results show that the YOLOv5l pre-trained on the COCO data set performs poorly on detecting victims, and the $AP@[.5 : .95]$ is only 19.5%. The model that uses our composite images as fine-tuning data can effectively detect victims, and increases the $AP@[.5 : .95]$ to 33.6%. The $AP@0.5$ increases from 32.4% to 53.4%. Our unsupervised harmonization method further improves the results by 2.1% and 6.1%, respectively.

1. INTRODUCTION

Object detection is an important topic that has been investigated for nearly 20 years. With the rise of deep learning (DL) and the availability of massive training data in recent years, DL-based object detection methods have made outstanding achievements and have become dominant (Zhao et al., 2019). Large data sets such as VOC (Everingham et al., 2015), ImageNet (Deng et al., 2009), and COCO (Lin et al., 2014) enable researchers to train a common object detector. There are also a lot of annotated data sets that make it possible to detect specific objects. For example, Wider Face (Yang et al., 2016) is a data set designed for face detection. Mappilary (Neuhold et al., 2017) provides 65 classes for object detection in autonomous driving scenes. Fruit and crop detection data sets are also available (David et al., 2020; Bargoti and Underwood, 2017). A high-accuracy detector relies heavily on a large number of training images, but in some special scenes training images are difficult to obtain. For instance, it is useful to train a victim detector that can be used by unmanned aerial vehicles (UAVs) in a rescue mission, but it is difficult to acquire such real images for training. Existing victim detection networks rely on common object data sets, which do not contain real victim images (Hoshino et al., 2021). Hartawan et al. (2019) trained a detector using INRIA person data set (Dalal and Triggs, 2005). The performance of these models on real victim images also needs more verification. Sulistijono and Risnumawan (2016) used only 19 real victim images to test their detector, which was not convincing.

As the number of images is a crucial factor in training good deep learning models, many researchers have studied how to generate synthetic data that can be used in training. Dwibedi et al. (2017) proposed a simple but effective way to augment data for instance detection of indoor objects. They collected some object instances and pasted them on random backgrounds to generate more training images. Wang et al. (2019) used a similar method that replaced an object instance with another instance of the same class.

Besides, it is convenient to use advanced computer graphics to generate a large number of realistically rendered images, which makes up for the lack of real data for training. In addition to many rendered data sets in the field of semantic segmentation (McCormac et al., 2017; Ros et al., 2016; Kirsanov et al., 2019; Zhang et al., 2021, 2022), researchers also rendered some synthetic data sets to train better object detection models. Han et al. (2021) proposed a rendered 3D face data set to study the relationships between object features and the performance of face detection. Peng et al. (2015) created 3D synthetic models to augment the training and outperformed previous methods. Rozantsev et al. (2015) proposed a method to synthesize unlimited unmanned aerial vehicles (UAVs) images in arbitrary 3D poses, and improved their UAV detector.

To facilitate first responders' rescue missions and save more lives we aim at training a victim detector to search for victims who are partially buried under ruins after an earthquake or building collapse. In general, when a person is crushed under the ruins, only part of the body is exposed, and the color of the body or clothes is similar to the background color due to dust or soil. A person detector trained on a normal object detection data set or a specific pedestrian data set might work, but the performance will likely be poor because these data sets

* Corresponding author

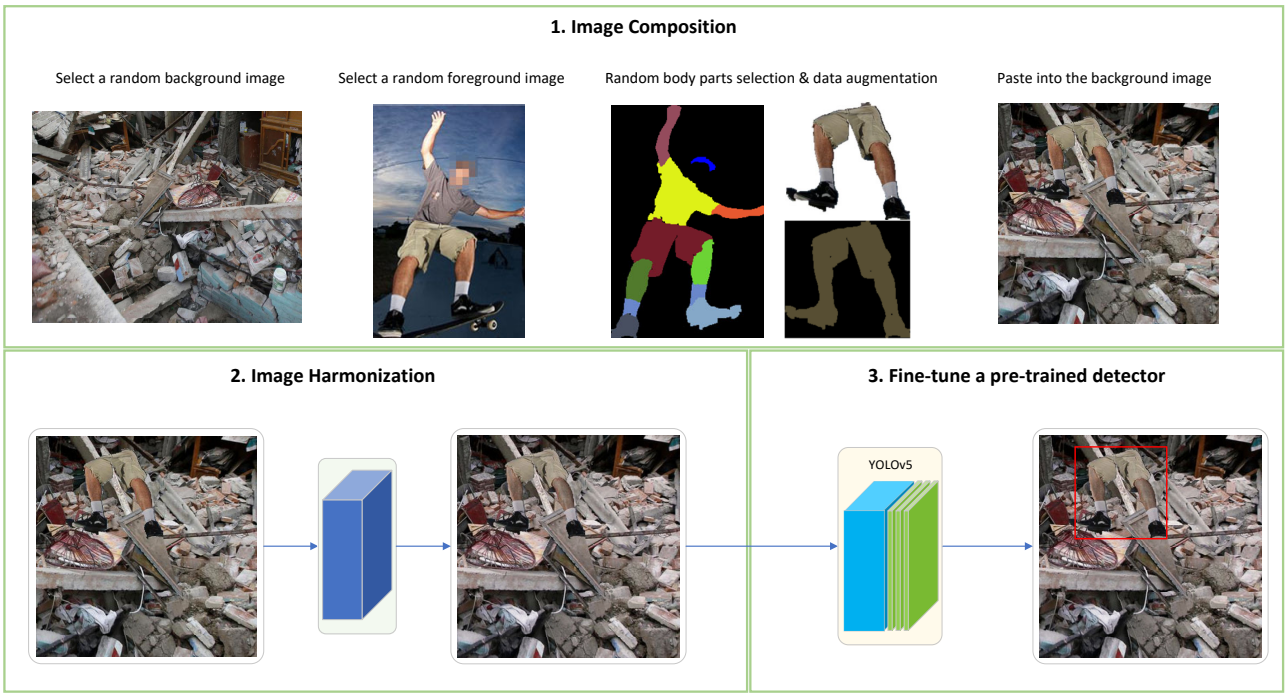

Figure 1. Our approach pipeline includes three steps: (i) image composition, (ii) image harmonization, and (iii) fine-tune a detector.

usually contain completely displayed, standing people in normal scenes. Therefore, in this paper we propose a composite data set for victim body part detection. We first randomly cut out human body parts from the open source human parsing data set LIP (Gong et al., 2017), and paste them into random background with collapsed structures. Then we use a novel unsupervised Generative Adversarial Network (GAN) to harmonize the body parts to fit the style of the background. Our contribution can be summarized as follows:

- We propose a novel framework to generate a data set that contains harmonious composite images of human body parts in ruins.
- We use the generated composite images to train a victim detector, and the experimental results show that our composite data is effective when training a victim detector. Our source code can be found on our project website https://github.com/noahzn/VictimDet.

Our approach pipeline is shown in Figure 1, which consists of three steps: image composition, image harmonization, and fine-tuning a victim detector. The rest of this paper is organized as follows. We present the image composition part in Section 2. Section 3 introduces the details of our deep harmonization network. Our experiments are elaborated in Section 4. Section 5 concludes the paper.

2. IMAGE COMPOSITION

The first step generates a composite image with simple cut and paste. Since we focus on victim detection in ruins, we need to collect both human body parts images as the foreground, and images with ruins as the background.

2.1 Collect background images

We use search keywords such as *earthquake*, *ruins* and *collapse* to collect background images I_b from Google images. We check to make sure there are no human beings in these images.

2.2 Collect foreground images

To obtain foreground images I^f we have two options. The first option is that, as used by Dwibedi et al. (2017) and Ghiasi et al. (2021), we can cut out complete human instances from existing data sets that contains the human class. However, in real scenarios it could happen that most of a person's body is buried, with only one arm or one leg exposed. Therefore, this option is not flexible enough to make composite images with only limbs exposed. The second option is to cut out a specific body part of a person as the foreground, and this option is better than the first one because in this case even if the detector only detects one arm or one leg, it can classify it as a potential victim. In this paper we also use the second option, and we start from an open-source human parsing data set LIP (Gong et al., 2017). LIP provides more than $50K$ human images annotated with 19 semantic classes such as *face*, *left arm*, *right arm*, *upper clothes*, *etc*. We can select and cut out specific semantic body parts from the foreground. The binary mask of the body parts can be denoted as M^f.

As shown in the first row of Figure 2, not all the images in the LIP data set are suitable for compositing victim images. Because there is no automatic method to accurately filter out all black-and-white images, blurred images, low-resolution images, severely occluded images and images with no exposed body parts, we manually deleted these images, and some good image samples are shown in the second row of Figure 2. These images have higher resolution, and the postures of the characters in the images are suitable to generate composite images in rescue scenes.

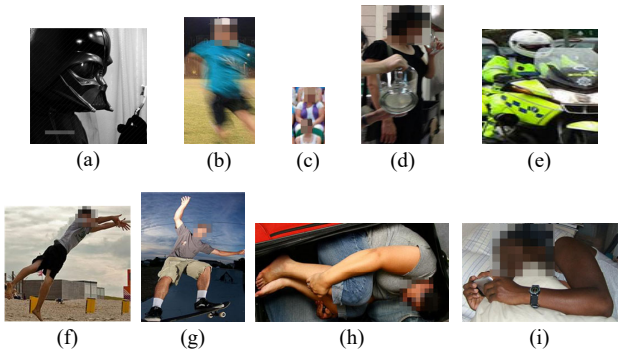

Figure 2. Some images in the LIP data set are not suitable to be used as the foreground, such as (a) black-and-white image; (b) blurred image; (c) low resolution image; (d) severe occlusion image, and (e) the image with no body parts exposed. (f)-(i) are good image samples we keep to generate composite images. We blur faces for privacy reasons.

2.3 Paste body parts into disaster scenes

For privacy reasons we discard faces in the images. At the same time we merge 19 semantic classes into five body parts combinations: *upper limbs*, *upper limbs + torso*, *lower limbs*, *lower limbs + torso*, and *full body*. For each image we randomly cut out body parts according to these five combinations. We also apply data augmentation such as *resize*, *crop*, and *flip horizontally*, to increase the diversity of foreground images. We paste the body parts at random positions onto the background images. The composite image I^c can be represented by the background I^b, the binary mask M^f, and the foreground I^f as:

$$I^c = I^b \times (1 - M^f) + I^f \times M^f. \tag{1}$$

3. IMAGE HARMONIZATION

Different from the tasks of Dwibedi et al. (2017) and Ghiasi et al. (2021) there are great visual differences between a foreground and a background in our task because of the inconsistent style (color, illumination, texture). Dwibedi et al. (2017) used Gaussian blending to smooth edges of the foreground, but this method cannot change the foreground's color, illumination, or texture. To make the composite images look more realistic we propose an unsupervised image harmonization network to adjust the style of the body part. Our proposed framework for image harmonization is based on the adversarial training. As shown in Figure 3 it consists of a generator G and two discriminators D_{global}, D_{local}. The generator generates a harmonious image, and two discriminators discriminate the real images and the generated harmonious images globally and locally, respectively. Using only one global discriminator will ignore the relationship between small-sized body parts and their surrounding background pixels, so we introduce a local discriminator to realize the harmonization of local illumination, texture and color characteristics.

3.1 Generator

The structure of our generator G is a U-Net with 3 attention layers. DoveNet (Cong et al., 2020) also uses the same generator structure, but we only use a three-channel composite image as input instead of using an extra mask channel, because we find that the combination of using an extra mask and our loss functions yields black artifacts on the output.

Given a composite image I^c we want the network to output a harmonious image I^h. The network should be trained to keep the background unchanged and make the foreground have the same style as the background, while the content does not change. For each pixel i, whose value is in the range of $[0, 1]$, we calculate the masked smooth L1 loss:

$$L_{1,i} = \begin{cases} \frac{1}{2}(I_i^c - I_i^h)^2 \times (1 - M_i^f) & \left|I_i^c - I_i^h\right| < 1, \\ (\left|I_i^c - I_i^h\right| - \frac{1}{2}) \times (1 - M_i^f) & otherwise. \end{cases} \tag{2}$$

Then, the loss of the whole image is:

$$L_1 = \sum_i^I L_{1,i}. \tag{3}$$

Compared with L1 loss smooth L1 loss avoids gradient explosion in some cases (Girshick, 2015).

3.2 Global discriminator

We use a global discriminator D_{global} to discriminate whether the input image is real or composite. Because we do not have the corresponding real version of a composite image, we take the background image I^b used to generate the composite image as the real image. The background image can be seen as a harmonious image, so the global discriminator can help the generator to generate harmonious images at a global level. D_{global} uses a PatchGAN (Isola et al., 2017) structure, and the adversarial loss function to train the global discriminator can be defined as:

$$\begin{aligned} L_{D_{global}} &= \mathbb{E}[\log D_{global}(I^b)] + \mathbb{E}[\log(1 - D_{global}(I^h))], \\ L_{G_{global}} &= \mathbb{E}[\log(1 - D_{global}(I^h))]. \end{aligned} \tag{4}$$

3.3 Local discriminator

We use another local discriminator D_{local} to focus on the foreground and its surrounding background, and constrain the local style consistency. The input is $\mathcal{P}(I^h)$, a patch centered on the foreground body parts and expands the neighborhood of the foreground bounding box by 60 pixels. So the patch contains both the foreground body part and its surrounding background pixels. The loss function is defined as:

$$\begin{aligned} L_{D_{local}} &= \mathbb{E}[\log D_{local}(\mathcal{P}(I^b))] + \mathbb{E}[\log(1 - D_{local}(\mathcal{P}(I^h)))], \\ L_{G_{local}} &= \mathbb{E}[\log(1 - D_{local}(\mathcal{P}(I^h)))], \end{aligned} \tag{5}$$

where $\mathcal{P}(I^b)$ is the corresponding patch on the background image. The local discriminator focuses on the cropped patches and is helpful for discriminating small foreground images.

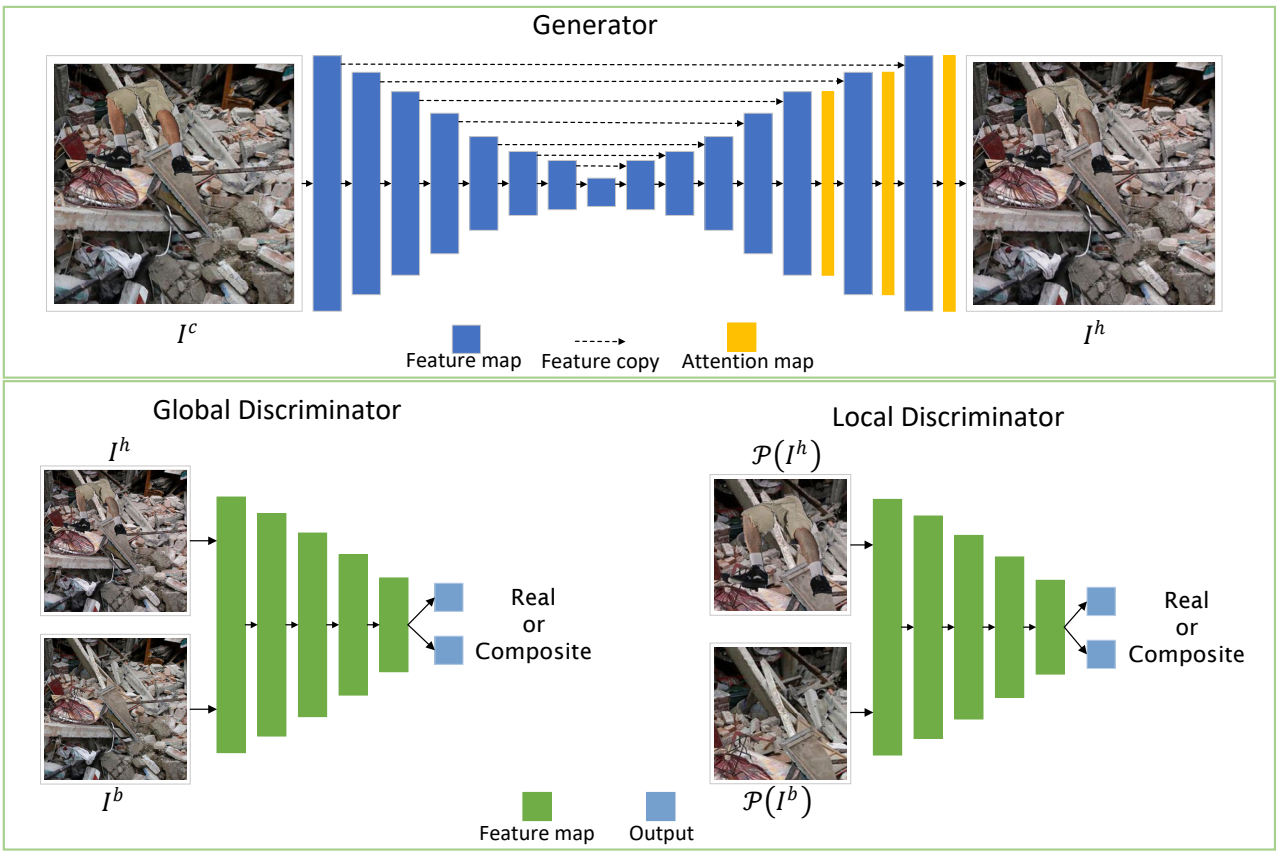

Figure 3. Our framework consists of a generator G and two five-layers discriminators D_{global}, D_{local}. The generator takes a composite image as input, and generates a harmonious image. Two discriminators discriminate the real images and the generated harmonious images globally and locally, respectively.

3.4 Locally constrained perceptual loss

The body parts in the output image should have the same semantic information as that in the input image. Perceptual loss proposed by Johnson et al. (2016) enforces the similarity between images at features level, and it has been used in many tasks (Rad et al., 2019; Yang et al., 2018; Ledig et al., 2017). The perceptual loss includes the content loss and the style loss. The content loss constrains the high-level semantic information, while the style loss makes two images consistent in style, such as color, illumination and texture. Different from those methods by computing the perceptual loss between the output image and the corresponding ground-truth, we propose to compute the perceptual loss between the input image and the output image. This is based on our purpose that the image harmonization network should only change the style of the input image, but not its semantic information. Besides, we compute the content loss on the same cropped patch $\mathcal{P}(I^h)$, as used in the local discriminator. The proposed locally constrained content loss L_{LCC} is defined as:

$$L_{LCC} = \frac{1}{C_j M_j N_j} \|\phi_j(\mathcal{P}(I^h)) - \phi_j(\mathcal{P}(I^c))\|_2^2, \quad (6)$$

where ϕ_j denotes the feature map of the j-th convolutional layer of a pretrained VGG16 model. $C_j \times M_j \times N_j$ is the size of the feature map, and $\|\cdot\|_2$ computes the l^2-norm. The shallow layers of a CNN model represent low-level style features such as colors and edges, and the deeper layers represent high-level semantic and content information (Lee and Tseng, 2019). We choose $j = 8, 11$ to compute the content loss.

Similarly, our style loss is also locally constrained and it only changes the style within the mask. It is defined as:

$$L_{LCS} = \frac{1}{C_j M_j N_j} \|(\mathcal{G}(\phi_j(\mathcal{P}(I^h))) - \mathcal{G}(\phi_j(\mathcal{P}(I^b))))\|_2^2, \quad (7)$$

where $\mathcal{G}$ is the Gram matrix proposed in the perceptual loss (Johnson et al., 2016). We choose $j = 3, 5$, which are two shallow layers to compute the style loss. We also use the total variation loss L_{TV} to smooth the local patch (Rudin et al., 1992), and it can be expressed as:

$$\begin{aligned} L_{TV} = \sum_{m,n} & \left| \mathcal{P}(I^h)_{m,n} - \mathcal{P}(I^h)_{m+1,n} \right| \\ & + \left| \mathcal{P}(I^h)_{m,n} - \mathcal{P}(I^h)_{m,n+1} \right|, \end{aligned} \quad (8)$$

where $\mathcal{P}(I^h)_{m,n}$ denotes the pixel value of the coordinates (m, n) on the patch $\mathcal{P}(I^h)$. The combined loss for training the generator G can be expressed as:

$$\begin{aligned} L_G = & \lambda_1 L_1 + \lambda_2 L_{LCC} + \lambda_3 L_{LCS} \\ & + \lambda_4 L_{G_{global}} + \lambda_5 L_{G_{local}} + \lambda_6 L_{TV}. \end{aligned} \quad (9)$$

We set $\lambda_1 = 80, \lambda_2 = 2, \lambda_3 = 0.2, \lambda_4 = 1, \lambda_5 = 1$ to make each loss part have a close scale. λ_6 is set to 10^{-5} for a slight regularization.

4. EXPERIMENTS

In this section we evaluate if the images generated by the proposed approach are helpful to train a victim detector. Our goal is not to improve the accuracy by directly enhancing the detector, but to fine-tune a pre-trained detector with additional composite images.

4.1 Data set and implementation details

We generate 1936 composite images, and harmonize them using the proposed framework. Each image has a size of 512×512 pixels, and we generate the ground-truth of bounding boxes of body parts. The purpose of generating these composite images is to use them to train a body parts detector. We also collect 197 real images to test if our generated images help to improve the accuracy of body parts detection. Some images are collected from the internet, and the main search keywords are earthquake rescue and collapse rescue. Most of the victims in these images are buried under debris, and we can only see part of their bodies. We also take some pictures by ourselves. All the test images are annotated. For fine-tuning and inference we use the official implementation of YOLOv5[1] on an Ubuntu 18.04 system with a Nvidia Titan XP graphics card.

4.2 Qualitative analysis of harmonized images

Figure 4 shows some samples of composite images (first row) and the corresponding harmonious version (second row) generated by our proposed method. The proposed unsupervised harmonization framework successfully transfer the illumination and colors of the background images to the foreground body parts. Further, the arms and legs are automatically added with some gray colors, making them look realistic, because in real images the body parts of victims are usually dirty due to dust or soil. However, we have no ground-truth to evaluate the quality of these generated harmonized images. We can only evaluate whether the fine-tuning of victim detectors benefits from these harmonized images. The quantitative evaluation is in Section 4.4.

4.3 Details of fine-tuning a body part detector

We fine-tune the detection model YOLOv5 (Jocher et al., 2021), which is a PyTorch implementation of the YOLOv4 (Bochkovskiy et al., 2020) pre-trained on the COCO data set (Lin et al., 2014). The YOLOv5 model consists of a backbone to extract features, a neck to concatenate features, and a head to predict the class and the bounding box. According to the difference of network depth and width, we used three different sizes of YOLOV5 models (Jocher et al., 2021), namely YOLOV5s, YOLOV5m, and YOLOV5l. Because the pre-trained YOLOv5 model has advantages in feature extraction, we fine-tune the pre-trained detector by fixing the backbone and updating the weights of the neck and the head. We want the model to be able to apply the feature extraction ability learned from the COCO data set to our composite data set. Two settings of data set are used, (i) the composite images without harmonization, and (ii) the harmonized composite images.

[1] https://github.com/ultralytics/yolov5

4.4 Evaluation of victim detection

According to the evaluation metric used in Pascal VOC challenge (Everingham et al., 2010) we measure the AP (average precision) and assume a successful detection if the predicted bounding box has an IoU (Intersection over Union) greater than a threshold 0.5 with the ground-truth, and denote this as $AP@0.5$. We also evaluate $AP@[0.5 : 0.95]$, which is a COCO metric and can be calculated by averaging AP over different IoU thresholds, from 0.5 to 0.95 with a step 0.05 (Lin et al., 2014). Table 1 shows the results of the COCO pre-trained YOLOv5 model and our models on the test set. We can find from the first row of each model that YOLOv5s is fast, but it performs poorly in detecting victims. YOLOv5s's shallow structures cannot learn a good representation from the training data. Although YOLOv5l is about 2.5 times slower than YOLOv5s, the detection accuracy improves substantially when the models are fine-tuned with our composite image (the second row of each model). Our harmonious images further improve the results (the third row of each model), which verifies the effectiveness of the proposed approach. Figure 5 visualizes some detection results. The default COCO pre-trained model (the second row) cannot detect victims as expected.

Because of the uncertainty of image copyright, we do not show the detection results of real victims in this paper. It is worth reporting that our models can detect some body parts, but fail in detecting some complete bodies. For example, in our test set there is a picture of the victim lying on the ground covered with mud, and the COCO pre-trained model cannot detect any body parts in this image. Although our method only detects one foot of the victim, not the whole body, it is still useful in the real rescue operation.

5. CONCLUSION

In order to enable first responders to find the victims partially buried in the ruins more efficiently in real rescue, we propose a novel framework to generate composite victims-in-ruins images and apply them to fine-tune a victim detector. The experimental results show that the normal COCO pre-trained models achieves low AP in detecting victims in ruins. Since it is difficult to get more real training images, our method uses composite images to train victim detectors, and its effectiveness is verified. We evaluate three variants of YOLOv5, which are fast detectors that can be deployed on UAVs for real-time victim detection. We hope that the work is useful in real disaster search and rescue and can save more lives.

There is still much room for improvement in studying the victim detection models, and our future work will focus on the use of UAVs with illumination for victim detection at night, because many rescues are carried out at night.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme and the Korean Government under Grant Agreement No 833435. Content reflects only the authors' view and the Research Executive Agency (REA) and the European Commission are not responsible for any use that may be made of the information it contains.

Figure 4. Qualitative results. The first row shows composite images, and the second row is the corresponding harmonious images generated by our network.

Table 1. Comparison of the accuracy of the YOLOv5 pretrained on COCO and the proposed fine-tuning models.

Model	Speed (ms)	Fine-tuned on composite	Fine-tuned on harmonious	AP@[0.5:0.95]	AP@0.5
YOLOv5s	4.7	✗	✗	11.6	21.2
		✓	✗	17.7	35.2
		✓	✓	18.1	35.3
YOLOv5m	8.1	✗	✗	16.6	28.9
		✓	✗	32.6	52.5
		✓	✓	34.3	55.2
YOLOv5l	11.2	✗	✗	19.5	32.4
		✓	✗	33.6	53.4
		✓	✓	35.7	59.5

References

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards. *ICRA*, 3626–3633.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y. M., 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L., 2020. Dovenet: Deep image harmonization via domain verification. *CVPR*, 8394–8403.

Dalal, N., Triggs, B., 2005. INRIA person dataset. *Online: http://pascal. inrialpes. fr/data/human*.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M. A. et al., 2020. Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *CVPR*, 248–255.

Dwibedi, D., Misra, I., Hebert, M., 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. *ICCV*, 1301–1310.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1), 98–136.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), 303–338.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., Zoph, B., 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. *CVPR*, 2918–2928.

Girshick, R., 2015. Fast r-cnn. *ICCV*, 1440–1448.

Figure 5. Visualization of victim detection. The first row is the ground-truth, and the second row is the default COCO pre-trained YOLOv5l model. The third row and the fourth row are the models fine-tuned on our composite images and harmonious images, respectively.

Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L., 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *CVPR*, 932–940.

Han, J., Karaoglu, S., Le, H.-A., Gevers, T., 2021. Object features and face detection performance: Analyses with 3d-rendered synthetic data. *ICPR*, 9959–9966.

Hartawan, D. R., Purboyo, T. W., Setianingsih, C., 2019. Disaster victims detection system using convolutional neural network (cnn) method. *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, IEEE, 105–111.

Hoshino, W., Seo, J., Yamazaki, Y., 2021. A study for detecting disaster victims using multi-copter drone with a thermographic camera and image object recognition by ssd. *2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, IEEE, 162–167.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *CVPR*, 1125–1134.

Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M. T., Marc, albinxavi, fatih, oleg, wanghaoyang0106, 2021. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support. doi:10.5281/zenodo.5563715.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 694–711.

Kirsanov, P., Gaskarov, A., Konokhov, F., Sofiiuk, K., Vorontsova, A., Slinko, I., Zhukov, D., Bykov, S., Barinova, O., Konushin, A., 2019. Discoman: Dataset of indoor scenes for odometry, mapping and navigation. *IROS*, 2470–2477.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, 4681–4690.

Lee, P.-Y., Tseng, C.-C., 2019. On the layer choice of the image style transfer using convolutional neural networks. *ICCE-TW*, 1–2.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *ECCV*, 740–755.

McCormac, J., Handa, A., Leutenegger, S., Davison, A. J., 2017. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? *ICCV*, 2678–2687.

Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. *ICCV*, 4990–4999.

Peng, X., Sun, B., Ali, K., Saenko, K., 2015. Learning deep object detectors from 3d models. *ICCV*, 1278–1286.

Rad, M. S., Bozorgtabar, B., Marti, U.-V., Basler, M., Ekenel, H. K., Thiran, J.-P., 2019. Srobb: Targeted perceptual loss for single image super-resolution. *ICCV*, 2710–2719.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A. M., 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. *CVPR*, 3234–3243.

Rozantsev, A., Lepetit, V., Fua, P., 2015. On rendering synthetic images for training an object detector. *CVIU*, 137, 24–37.

Rudin, L. I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4), 259–268.

Sulistijono, I. A., Risnumawan, A., 2016. From concrete to abstract: Multilayer neural networks for disaster victims detection. *2016 International electronics symposium (IES)*, IEEE, 93–98.

Wang, H., Wang, Q., Yang, F., Zhang, W., Zuo, W., 2019. Data augmentation for object detection via progressive and selective instance-switching. *arXiv preprint arXiv:1906.00358*.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., Wang, G., 2018. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE TMI*, 37(6), 1348–1357.

Yang, S., Luo, P., Loy, C.-C., Tang, X., 2016. Wider face: A face detection benchmark. *CVPR*, 5525–5533.

Zhang, N., Nex, F., Kerle, N., Vosselman, G., 2021. Towards Learning Low-Light Indoor Semantic Segmentation with Illumination-Invariant Features. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 427–432.

Zhang, N., Nex, F., Kerle, N., Vosselman, G., 2022. LISU: Low-light indoor scene understanding with joint learning of reflectance restoration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 470–481.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., Wu, X., 2019. Object detection with deep learning: A review. *IEEE TNNLS*, 30(11), 3212–3232.