Detecting Nonexistent Pedestrians

Jui-Ting Chien, Chia-Jung Chou, Ding-Jie Chen, Hwann-Tzong Chen
National Tsing Hua University, Taiwan
{ydnaandy123,jessie33321,djchen.tw}@gmail.com, htchen@cs.nthu.edu.tw

Abstract

We explore beyond object detection and semantic segmentation, and propose to address the problem of estimating the presence probabilities of nonexistent pedestrians in a street scene. Our method builds upon a combination of generative and discriminative procedures to achieve the perceptual capability of figuring out missing visual information. We adopt state-of-the-art inpainting techniques to generate the training data for nonexistent pedestrian detection. The learned detector can predict the probability of observing a pedestrian at some location in image, even if that location exhibits only the background. We evaluate our method by inserting pedestrians into images according to the presence probabilities and conducting user study to determine the 'realisticness' of synthetic images. The empirical results show that our method can capture the idea of where the reasonable places are for pedestrians to walk or stand in a street scene.

1. Introduction

Humans are good at inferring implicit information from images based on the context. For instance, with sufficient familiarity of urban street scenes, humans know where to look at to find pedestrians. It implies that peripheral cues in the scene other than the pedestrians themselves are useful for pedestrian detection. The ability to explore and exploit those cues would be important for algorithms to achieve human-level scene understanding and behavior analysis.

Our goal is to address the problem of predicting where the pedestrians are likely to appear in a scene. Such a task is new and different from conventional pedestrian detection in that the target locations might not contain any pedestrians. Many urban street scene datasets [4, 6, 7, 8] are available for training deep models to solve object detection and semantic segmentation, but they cannot be directly used for our task. We have to generate new training data that enforce our method to learn how to infer possible presence locations of pedestrians using implicit contextual cues of scenes rather than the features on the pedestrians.

Figure 1. Top: The input image and the predicted heatmap. The likelihoods of head and feet positions are depicted in red and blue. Bottom: The synthesized images with phantom pedestrians using the baseline method (left) and the proposed method (right).

We use adversarial training [10] and build a ConvNet that combines generative and discriminative procedures to achieve the perceptual capability of figuring out missing visual information. The contributions of this work include

1. Suggesting an alternative direction toward scene understanding, which is relatively unexplored and may be applied to other pedestrian- or scene-related tasks, *e.g.*, as auxiliary information for pedestrian accident avoidance.

2. Providing a training dataset for the new task of predicting pedestrian presences.

3. Proposing a ConvNet that incorporates an adversarial loss. The results show that our model successfully predicts the probability of observing a pedestrian at some location in the current image, even if that location exhibits only the background, as shown at the top row in Fig. 1.

4. Designing a pipeline for automatic image synthesis, which is used to evaluate our method and can be further used as auto-annotated training data. Examples of synthesized images are shown at the bottom row in Fig. 1.

2. Related work

The idea of *Binge Watching* proposed by Wang *et al*. [26], which is to predict affordance poses, is very similar to our idea of nonexistent pedestrian detection. Given a new scene, the *variational autoencoder* (VAE) can be used to sample from the distribution of possible poses. Wang *et al*. successfully show the possibility of predicting reasonable nonexistent human poses using ConvNets. However, their data are collected only from some specific indoor scenes of seven different sitcoms. Our work focuses more on outdoor street scenes and can generate training data with various combinations. Huang and Ramanan [12] generate synthetic pedestrian images by inserting CG rendered human models into real street scene images, and use GAN to make the generated images more realistic. In contrast, we use real 2D human images for synthesizing. We also use the predicted heatmap as the prior of reasonable places for pedestrians. Sun and Jacobs [24] aim to predict where missing curb ramps should exist in an image, even when no curb ramps are present. They generate training data by covering curb ramps with black masks. Our training data, on the other hand, are generated via removing the pedestrians, using state-of-the-art inpainting techniques to reduce the artifact features.

Deep convolutional neural networks. Deep ConvNets are popular in computer vision. Many challenging problems (*e.g*., image classification [15], object detection [9], human pose estimation [25], *etc*.) can be solved via end-to-end learning with ConvNes. *Fully convolutional networks* (FCN) [23] replaces fully connected layers in the network with convolutional layers and produces an output whose size is the same as the input size. The significant improvement of segmentation accuracy makes FCN be widely used in pixel-wise prediction problems. The stacked hourglass networks [18] further use the concept of residual maps and becomes the state-of-the-art method for human pose heatmap prediction. Our work is based on the two aforementioned networks [18, 23] to develop a more suitable architecture for our task.

Generative adversarial networks. GAN [10] consists of two components: a generator (G) and a discriminator (D). The training is unsupervisedly done by playing a two-player minimax game (*i.e*. adversarial game) with the objective function $V(G, D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z))] \,, \tag{1}$$

where D aims to distinguish two data domains and G tries to map a random normal distribution z to a target data space p (*e.g*., a set of human face images). DCGAN [22] provides a more stable and easier-to-train framework by replacing pooling layers with strided convolution. WGAN [2] uses Wasserstein distance to measure the loss and successfully avoids mode collapsing in various generator architectures. WGANGP [11], based on WGAN, improves the way how Lipschitz constraint is enforced by replacing weight-clipping with gradient penalty. CGAN [17] extends GAN to supervised learning with additional constrains. Many related applications use the concept of GAN to achieve further improvements, such as pose estimation [3], saliency detection [20, 19], semantic segmentation [16, 27], and image inpainting [13, 21, 28]. We apply GAN to our task of detecting nonexistent pedestrians.

Figure 2. What is a nonexistent pedestrian? Left column: Images of street scenes. Right column: Everything stays the same except that some pedestrians disappear. The white regions are places that have once existed some pedestrians, and now they are considered reasonable places for pedestrians to walk or stand.

3. Approach

The proposed pipeline has three parts: *i*) generating training data for nonexistent-pedestrian detection, *ii*) learning to predict possible locations of pedestrians, and *iii*) synthesizing new images with phantom pedestrians for qualitative evaluation.

3.1. Training Data

To address our task, we make the following assumption: if a place observed in the real world does exist pedestrians, then that place is likely to be a suitable one for pedestrians to appear. Fig. 2 gives some practical examples showing what our hypothesis looks like. Under this assumption, we automatically generate our training dataset by collecting images, estimating pedestrian poses, and removing pedestrians. Fig. 3 shows some examples of the training pairs of input image and ground truth heatmap (as output) used in our experiments.

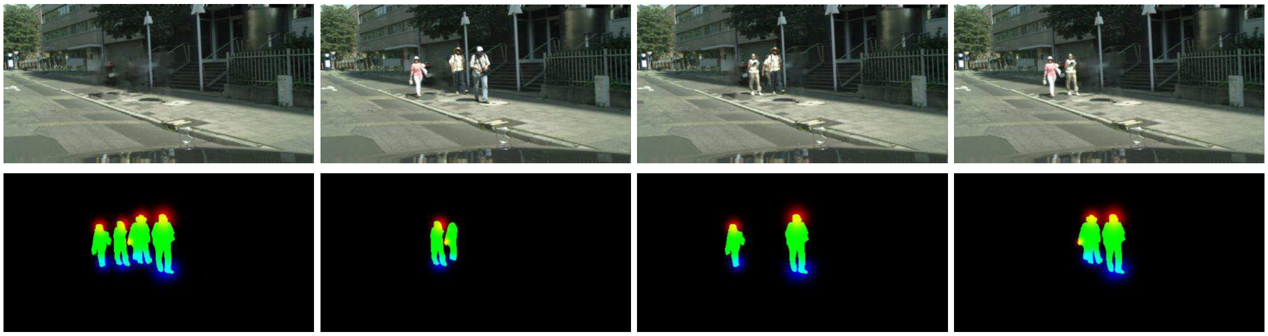
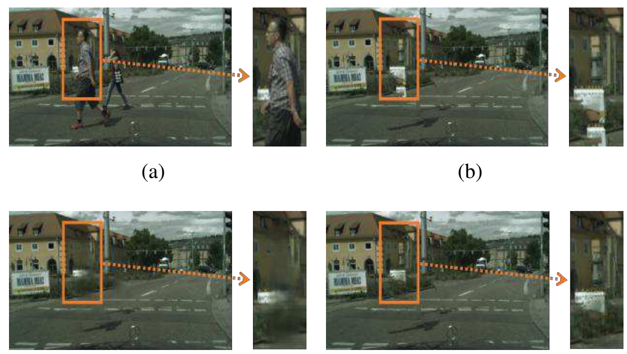


Figure 3. Training data pairs. Top row: Input images with pedestrians removed randomly. Bottom row: Corresponding heatmaps as the ground truths. Blue, red, and green color represent the likelihoods of head, foot, and full body respectively.

Collecting images. We collect a wide range of urban-scene images from the *Cityscapes* dataset [4]. It is a large-scale dataset of street scenes in 50 different European cities. It provides 2,975 training images and the corresponding pixel-level annotations. However, more than 2,500 images do not contain any pedestrians and hence a large number of pixels are labeled as non-pedestrian. The unbalanced pixel amounts between pedestrian and non-pedestrian might lead to bias, and the network would tend to generate all-empty predictions. To alleviate the bias, we only select those images of which the pedestrian pixels cover more than 5% of the entire image. Furthermore, for the sake of reducing time and space, we resize the images to 256×512 in the subsequent steps.

Estimating pedestrian poses. We further enrich our training data information by analyzing pedestrian poses. The pedestrian poses are estimated using the stacked hourglass network [18]. We train the stacked hourglass on the MPII [1] and LSP [14] datasets. During inference, we crop out each single pedestrian from the original scene to get the predicted pose. We superimpose the predicted locations of head and feet onto a blank image of the same size as the input image, and blur the locations to generate a 'ground truth' heatmap of the original street scene. As shown in the bottom row of Fig. 3, besides the green color representing the segment of pedestrians, the blue and red channels indicate the possible positions for head and feet respectively.

Removing pedestrians. Since each image in the Cityscapes dataset has the ground truth segmentation of pedestrians, we can use state-of-the-art inpainting methods [5, 21, 28] to remove all pedestrians and create a set of 'synthetic background' images. The comparison of the three methods is shown in Fig. 4. The method of Criminisi *et al.* [5] is an exemplar-based method which repeatedly sample small patches from original images to fill the holes. The results preserve the original image texture but



(a) (b) (c) (d)

Figure 4. Comparisons between different inpainting methods. (a) An original street view image. (b) An exemplar-base approach by Criminisi *et al.* [5] (c) A variational autoencoder based approach by Pathak *et al.* [21] (d) Image content and texture optimized jointly by Yang *et al.* [28]

the content might not make sense as shown in Fig. 4(b). Pathak *et al.* [21] train a variational auto-encoder (VAE) combined with a discriminator to render realistic images that can deceive the discriminator. The generated images have reasonable image content and structure, but tend to be blurry as shown in Fig. 4(c). Yang *et al.* [28] integrate the advantages of previous methods by optimizing the texture and content of the image jointly. They fill the hole by selecting the patch that not only makes the target region boundary seamless but also maintains the structure of scene. The results preserve high-resolution details and have reasonable content structure as shown in Fig. 4(d).

Producing training images. Based on the synthetic backgrounds, the original images, and the ground truth segmentations, we generate various training data exhibiting different combinations of removed/non-removed pedestrians and the corresponding heatmaps of possible pedestrian presences as shown in Fig. 3. By doing so, we can acquire train-

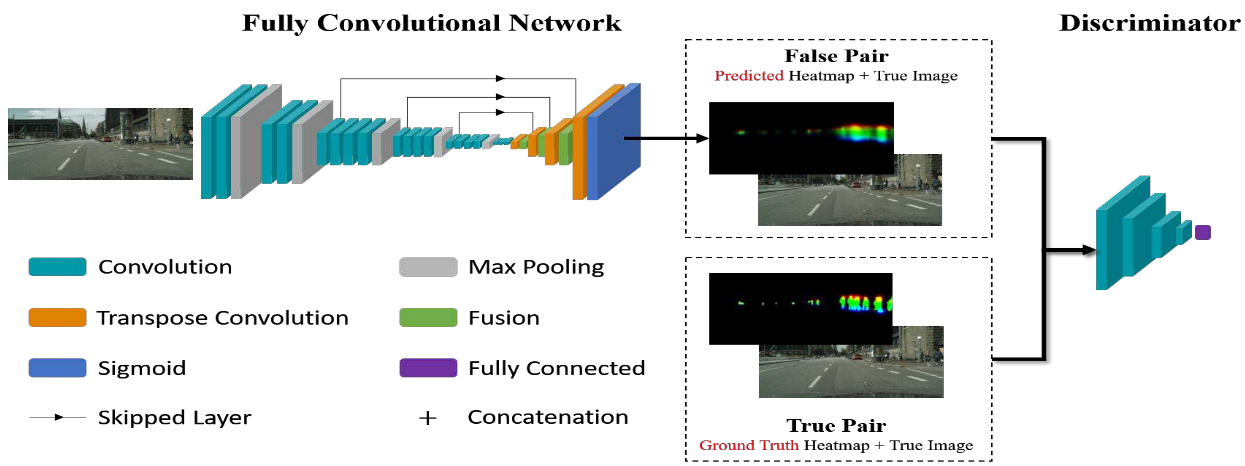


Figure 5. The architecture of the proposed model FCN+D.

ing data for learning relations not only between pedestrian and scene structure, but also among pedestrians, *e.g.*, how pedestrians interact with each other. We extract individual pedestrian according to the instance-level annotation provided by [4]. For each combination, the selected pedestrians are added to the synthetic backgrounds (top row of Fig. 3), and unselected pedestrian heatmap segmentations are added to the ground truth heatmap (bottom row of Fig. 3). We only produce at most 40 images of different combinations from one single synthetic background.

3.2. Learning

We formulate the problem as a multi-label classification problem: The input is an RGB image, and the output is a 3-channel heatmap which is for head, feet, and full body, respectively. Each class is independent and not mutually exclusive (*i.e.*, a pixel can be head and body at the same time). Some examples of the pairs of training image and ground truth are shown in Fig. 3.

Network architecture. We present three convolutional networks to detect nonexistent pedestrians. We adopt the network of FCN [23] and the *stacked hourglass network* [18] as our first and second models. Our FCN is based on VGG-19 and our stacked hourglass network is a two-stack version. We also propose a new model by combining FCN with a discriminator (D). Our discriminator network architecture is adapted from DCGAN [22]. While the discriminator tries to judge whether the heatmap is generated by FCN or is picked from the ground truth set, FCN attempts to deceive the discriminator by generating heatmaps that look like the ground truths as much as possible. We also require the generated heatmap to be coherent with the input image. This constraint resembles conditional GAN [17] where our condition is the input image. We directly concatenate the input image with the heatmap generated by FCN as a 'fake' pair, and concatenate the input image with the corresponding ground truth as a 'true' pair. We refer to our proposed model as FCN+D, shown in Fig. 5.

Content loss. FCN [18] and the stacked hourglass network [23] produce heatmaps based on the content of input images. We minimize the per-pixel differences between the ground truth and the output heatmap. The content loss (L_{CE}) is defined by

$$L_{\text{CE}} = -\sum_{i=1}^{P}\sum_{c=1}^{C}\left[y_i^c \log \sigma(x_i^c) + (1-y_i^c)\log(1-\sigma(x_i^c))\right] , \tag{2}$$

where C is the number of classes, P is the number of pixels, x is the output heatmap, y is the ground truth, and $\sigma(\cdot)$ is the sigmoid function.

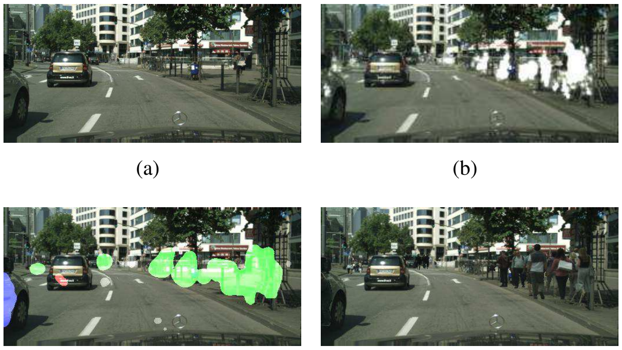
Adversarial loss. The proposed FCN+D uses an additional loss for adversarial training, where FCN is referred to as a generator (G). Our adversarial loss follows WGAN [2] using the Wasserstein distance. We use gradient penalty [11] to enforce the Lipschitz constraint. The adversarial losses for G and D are defined by

$$\begin{aligned} \mathcal{L}_{\text{advD}} &= -\mathbb{E}_{y\sim P_r}[D(y,I)] + \mathbb{E}_{x\sim P_g}[D(x,I))] \\ &\quad + \lambda_{gp}\, \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x},I)\|^2 - 1)^2] \end{aligned} \tag{3}$$

and

$$\mathcal{L}_{\text{advG}} = -\mathbb{E}_{x\sim P_g}[D(x,I))] , \tag{4}$$

where I is the input image, x is the generated heatmap, y is the corresponding ground truth heatmap, P_r is the real data distribution, and P_g is the model distribution implicitly



(a) (b) (c) (d)

Figure 6. The synthesis pipeline. (a) An arbitrary input street scene image. (b) The predicted nonexistent-pedestrian heatmap. (c) Synthesis guiding map based on the predicted nonexistent-pedestrian heatmap and the user preferences. (d) The final synthetic image.

defined by $x = G(I)$. We combine the content loss with the adversarial loss as the overall generator loss:

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{CE}} + \lambda_G \mathcal{L}_{\mathrm{advG}} , \tag{5}$$

where λ_G is a hyperparameter controlling the ratio. In this case, generator has an auxiliary adversarial loss that guides the direction of updating. During training, we alternately update the discriminator and the generator of the proposed FCN+D.

3.3. Synthesis

Our method automatically synthesizes a new test image according to the predicted nonexistent-pedestrian heatmap. The pipeline is shown in Fig. 6. Given an arbitrary street scene, we predict the nonexistent-pedestrian heatmap using the proposed model. Then, we produce another 'synthesis guiding map' based on the predicted full body heatmap and the user preferences. Finally, our method searches the most compatible and suitable pedestrians from the pedestrian dataset and produces the final synthetic output image.

Synthesis guiding map. We produce a 'synthesis guiding map' based on the predicted full body heatmap and the user preferences. An example is shown in Fig. 6(c). The synthesis guiding map is used as an interface between the user preferences and the final synthetic image. Since each pixel in the original predicted heatmap simply represents a likelihood, there are two questions to be answered by the user: *i*) how large the likelihood response should be for a pixel to be qualified as belonging to a pedestrian? and *ii*) how likely the pixels may connect to each other to form a shape of pedestrian? We provide two hyperparameters: Gaussian filter standard deviation σ and heatmap threshold α for users to adjust, as shown in Fig. 7. A Gaussian filter is used to

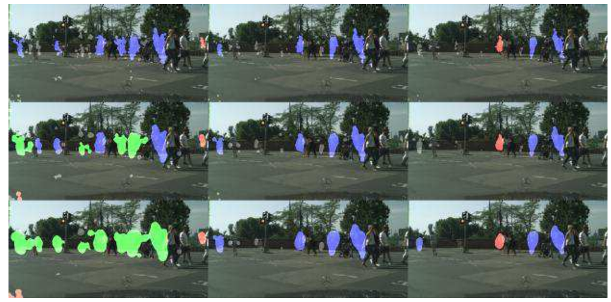


Figure 7. Various synthesis guiding maps for different user preferences. From top to bottom: smaller σ to larger σ. From left to right: smaller α to larger $alpha$. The color regions mean that the pixel responses are higher than the specified threshold α. White: local regions of which the responses are too small and can be ignored. Blue: local regions of which the aspect ratio is classified as a single pedestrian. Green: local regions of which the aspect ratio is classified as multiple pedestrians. Red: local regions that are classified as invalid regions (*e.g.*, floating in the air).

smooth the responses. A larger value of σ means that more pixels will be connected together to form a larger region instead of many separate local regions. After Gaussian filtering, we only consider those responses that are higher than α as the candidate places of pedestrians.

Pedestrian dataset. We collect images of pedestrians from [4, 7] to form our pedestrian dataset. Each pedestrian is categorized by height, width, and aspect ratio based on the segment bounding box. Besides 'single-person' images, we also collect 'multi-person' images from [4]. If people in an image overlap each other and their segments are not separable, then we crop the whole connected component and mark it as multi-person. We further estimate the pedestrian poses using the stacked hourglass network [18] for a more detailed representation. Currently we only use the keypoints of head and feet.

Rendering. For each test image, we obtain three heatmaps modeling the head, feet and body locations of pedestrians. We perform non-maximum suppression on the synthesis guiding map and propose several candidate 'phantom' pedestrians. After filtering out some unlikely or invalid phantoms (*e.g.*, too small or floating in the air), we classify the phantoms as a single person or multiple people by the aspect ratio of their bounding box. Finally, our method searches for the most likely real pedestrian(s) in the pedestrian dataset according to the poses and shapes, and the output image can be rendered by copy-and-paste.

4. Experiments

We conduct several experiments to evaluate our method on two aspects: the capability of nonexistent pedestrian detection and the quality of synthetic images.

4.1. Nonexistent Pedestrian Detection

Since our task is rather new and abstract, there is no standard method to evaluate the capability of detecting nonexistent pedestrians. We use the recall rate as a metric to give a rough comparison between the three models mentioned in Sec. 3.2. We also look into the predicted heatmap results and give some preliminary analysis.

Experiment settings. We use the dataset mentioned in Sec. 3.1 to train our three ConvNets: FCN, *stacked hourglass network*, and FCN+D. The training data contain around 12,000 images. The sizes of the training images and the ground truth labels are both 256×512. The earlier layers of FCN and the generator of FCN+D are initialized with the VGG-19 weights. The FCN+D loss function is set with $\lambda_{gp} = 10$ and $\lambda_G = 0.001$. We alternately update the discriminator and the generator of FCN+D. The batch size is 9 for FCN and FCN+D, and is 2 for the stacked hourglass network. We keep other settings and hyperparameters the same as in the original papers [18, 22, 23].

Validation. We randomly select nearly 700 different images generated from 10 synthetic backgrounds for validation. To evaluate quantitatively the performance of our networks on predicting heatmaps, we use the recall rate as the metric:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\text{area}(l_i \cap h_i)}{\text{area}(l_i)}\,, \tag{6}$$

where N is the number of images in the validation set, and $\text{area}(l_i \cap h_i)$ is the overlap between the ground truth heatmap l_i and the predicted whole-body heatmap h_i. We choose the recall rate as the evaluation metric because our ground truths just represent parts of many feasible solutions. We expect the network to have its own perceptual capability of figuring out missing visual information rather than being restricted with the ground truths provided by ourselves. The validation scores are shown in Table. 1. Intuitively, FCN+D seems to perform the best on retrieving those places overlapped with the ground truth. However, the recall rate should not be over-interpreted as our proposed architecture being always better in all respects. The results are more likely to reflect the facts that all of the models are similar on detecting what we call nonexistent pedestrians. No matter what architectures or loss functions are chosen, the detection results show no noticeable differences.

	FCN [23]	Hourglass [18]	FCN+D
Recall	0.86	0.88	0.89

Table 1. Comparison on the recall rate for the three methods.

Figure 8. Validation results. Left: Ground truth. Right: Predicted heatmap of full body. The responses almost appear only on the pedestrian-removed regions and many clear contours of pedestrians can be found during validation.

Testing. We test the proposed method using unseen street scene images, which are natural and do not contain artifacts. If our models only have the perceptual capability of figuring out inpainting artifacts, then there should not be any responses on these input images. It turns out that our ConvNet is still able to predict many reasonable places for pedestrians as shown in Fig. 9 during testing. The intensities of responses are smaller than those predicted during validation, but after we normalize the intensities the heatmap shows visible and reasonable responses. We think that our models actually learn both perceptual capabilities of finding inpainting artifacts and detecting nonexistent pedestrians.

Analysis. Given unseen images, the trained model performs well on mimicking human perceptual ability. In Fig. 9, we highlight some logical patterns that can be observed in the generated images: *i*) Sidewalks, safety islands, and bus stops are often assigned with high probabilities of pedestrian presence, even if the scene is void of pedestrians. *ii*) The timing is right: The 'phantom' pedestrians are inclined to cross the street when there is no car. *iii*) People tend to form groups: The 'phantom' like to stand or walk close to existing people. *iv*) Depth and perspective are correct: The 'sizes' of high-response areas in the heatmap are in accordance with the depth and vanishing point.

4.2. Synthetic Images with Rendered Pedestrians

In Sec. 3.3 we provide a complete process to automatically generate the synthetic output images. To verify the quality of the synthesized images, we conduct user study and ask the viewers to choose the most realistic image. Fig. 1 shows a simple pipeline. Given an input image and the corresponding predicted heatmap, we synthesize a new image with baselines and our proposed method, as shown at the bottom row of Fig. 1.

Figure 9. Some observed heatmap patterns. From top to bottom: appropriate positions, appropriate timing, forming groups, and correct perspective.

.

Baselines. We consider two baselines for comparison. The first one is to paste the same number of pedestrians into the target image as our proposed method. The pedestrian poses and positions are unconstrained and random. As shown in Fig. 10(a), the results look weired and unnatural. The second baseline directly copies the whole group of pedestrians from another image and paste them, under the same scale, into the target image. As shown in Fig. 10(b), the results are more realistic because the relations between the copied pedestrians are natural and correct. If the original scene structure is also similar to the target image structure, the output image may look appealing. Therefore, this baseline is actually quite competitive.

Experimental results. We recruit several viewers to participate in the experiment. To each viewer we ask 100 questions. Each question shows three synthetic images at the same time and the viewer has to choose the one that is considered the most realistic among the three. We count the votes of each method; getting more votes means better quality. Our method significantly outperforms the other two baselines. The experimental results show that the predicted positions are reasonable and our synthetic images are realistic. More examples of our synthesis pipeline and results are shown in Fig. 11.

(a)

(b)

Figure 10. Two baselines. (a) Pedestrians are randomly selected and placed. (b) Pedestrians are copied as a whole from another street scene image in the pedestrian dataset.

5. Conclusion and Future Work

This work presents a possibility of using deep networks to infer a scene from the implicit cues for nonexistent pedestrian detection, which is a relatively unexplored area and has no standard mechanisms for quantitative evaluation. The empirical results show that our method can capture the idea of where the reasonable places are for pedestrians to walk or stand in a street scene. The proposed idea suggests an alternative route toward scene understanding and may be further applied to other tasks, such as pedestrian segmentation, activity analysis, *etc*. Extending this work from image to video will be our future work. We also look for the improvement of synthetic image quality. Our current copy-and-paste method sometimes cannot achieve visually appealing quality. Using GANs to generate end-to-end synthetic output images would be an interesting direction.

References

[1] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3686–3693, 2014. 3

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017. 2, 4

[3] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *CoRR*, abs/1705.00389, 2017. 2

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223, 2016. 1, 3, 4, 5

[5] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Processing*, 13(9):1200–1212, 2004. 3

Figure 11. More examples of the synthesis pipeline. We adjust the image brightness for better visualization. From top to bottom: input images, predicted heatmaps, and synthesized images with phantom pedestrians according to the predicted heatmaps.

[6] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012. 1

[7] F. Flohr and D. Gavrila. Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues. In *British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013*, 2013. 1, 5

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *I. J. Robotics Res.*, 32(11):1231–1237, 2013. 1

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014. 2

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. 1, 2

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. 2, 4

[12] S. Huang and D. Ramanan. Recognition in-the-tail: Training detectors for unusual pedestrians with synthetic imposters. *CoRR*, abs/1703.06283, 2017. 2

[13] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, 36(4):107:1–107:14, 2017. 2

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, pages 1–11, 2010. 3

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'25*, pages 1106–1114, 2012. 2

[16] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *CoRR*, abs/1611.08408, 2016. 2

[17] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2, 4

[18] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499, 2016. 2, 3, 4, 5, 6

[19] H. Pan and H. Jiang. Supervised adversarial networks for image saliency detection. *CoRR*, abs/1704.07242, 2017. 2

[20] J. Pan, C. Canton-Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giró i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, abs/1701.01081, 2017. 2

[21] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544, 2016. 2, 3

[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 2, 4, 6

[23] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 2, 4, 6

[24] J. Sun and D. W. Jacobs. Seeing what is not there: Learning context to determine where objects are missing. *CoRR*, abs/1702.07971, 2017. 2

[25] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS'27*, pages 1799–1807, 2014. 2

[26] X. Wang, R. Girdhar, and A. Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR 2017*, 2017. 2

[27] Y. Xue, T. Xu, H. Zhang, R. Long, and X. Huang. SegAN: Adversarial Network with Multi-scale L_1 Loss for Medical Image Segmentation. *ArXiv e-prints*, June 2017. 2

[28] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. *CoRR*, abs/1611.09969, 2016. 2, 3