Cross-Domain Hallucination Network for Fine-Grained Object Recognition

Jin-Fu Lin[1], Yen-Liang Lin[2], Erh-Kan King[1], Hung-Ting Su[1], and Winston H. Hsu[1]

[1]National Taiwan University
[2]GE Global Research, Niskayuna, NY, USA
{*zendo3464, yenlianglintw, goan15910*}*@gmail.com*, {*d06944009, whsu*}*@ntu.edu.tw*

Abstract

Existing fine-grained object recognition methods often require high-resolution images to better discriminate the subordinate classes. However, this assumption does not always hold in current surveillance systems, where the distinguished parts may not be clearly presented. Besides, data insufficiency and class imbalance make the problem even more challenging. In this paper, we leverage high-resolution images collected from Internet to improve the vehicle recognition in the surveillance environments. A cross-domain hallucination network is proposed to minimize the domain discrepancy and enhance the quality of low-resolution surveillance images. To better align the cross-domain features and boost the recognition performance, we extend the original framework to part-based hallucination networks, where the parts are automatically extracted based on the maximum responses from the convolution filters. We evaluate our method on a public surveillance vehicle dataset (BoxCars21k). Experimental results demonstrate that our approach outperforms the state-of-the-art methods.

Keywords: cross-domain, hallucination, fine-grained classification

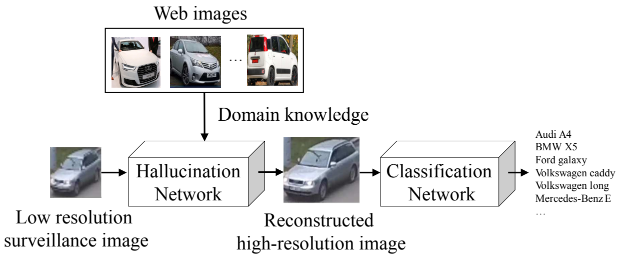

Figure 1. We propose a cross-domain hallucination network to transfer the knowledge from web to surveillance domain and improve the fine-grained object recognition in the low-resolution surveillance environments.

1. Introduction

High-level understanding and analysis of surveillance images enable numerous applications, for example, [5] predict the demographic attributes (e.g., income, per capita carbon emission, crime rates) by using the detected vehicles. However, unlike web images, surveillance images are often low resolution, which causes the important information loss (i.e., distinguished parts are not always visible). Moreover, it is time consuming to collect a large number of vehicle models and makes from the surveillance videos. Severe class imbalance also causes the additional difficulties to machine learning algorithms. There are abundant vehicle images on the Internet, which are often high resolution and the class distribution is relatively balanced. Web images possess richer domain knowledge and are thus beneficial to the recognition tasks. It would be interesting to see how to leverage the domain knowledge from additional high-resolution web data to improve the object recognition in the low-resolution surveillance domain, which significantly reduces the annotation efforts.

In this paper, we aim at improving fine-grained object recognition under the surveillance environments. To our best knowledge, there is only little research [20] that work on this novel problem. We observe that the recognition performance for fine-grained object recognition in surveillance images is highly affected by the quality of the input images. Even using all the label information from the surveillance domain, we still can not achieve satisfactory performance. Moreover, information loss of low resolution input images causes the features less representative, which makes it difficult to align the cross-domain feature distributions. The goal of this work is to improve the quality of input images by transferring the domain knowledge from other high resolution vehicle datasets.

The core idea of our method is illustrated in Figure

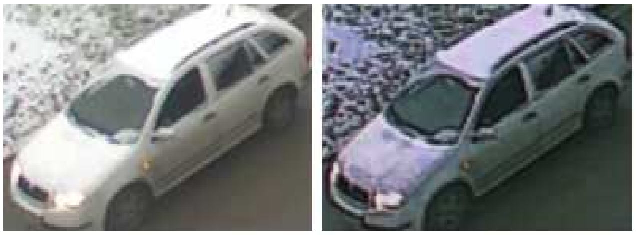

Figure 2. **Left image:** low resolution input image. **Right image:** image generated by our hallucination network.

1. We propose a cross-domain hallucination network that leverages high-resolution web images to improve the fine-grained vehicle classification of the low resolution surveillance images. The hallucination network minimizes the discrepancy between the two different domains and enhance the quality of the low-resolution images (see the right image in Figure 2). This transformation makes the surveillance images become more discriminative and facilitate the classification network to obtain more accurate results.

Motivated by [7], some parts of vehicles are informative and could be helpful for recognition. Thus, we further extend the framework to part-based hallucination networks. We automatically obtain the part regions of vehicle images for both domains by using the maximum responses of the feature maps in the convolution layers, which corresponds to the parts of a vehicle image (e.g., front, rear and roof). A part-based hallucination network is trained to minimize the domain differences on those part regions, where all part regions share one hallucination network. Finally, we ensemble two hallucination and classification networks (one for whole image and another for part regions) to obtain the final classification results.

The main contributions of this work are summarized as following:

- We propose a cross-domain hallucination network that enhances the discriminative details of the low resolution surveillance images by leveraging the additional high-resolution web images.
- We investigate a two-step classification network that experimentally outperforms several state-of-the-art methods in a public surveillance vehicle dataset.
- We explore part-based hallucination networks, and ensemble the whole and part-based hallucination networks to further boost the performance.

2. Related work

Vehicle plate recognition: previous work [15, 3, 1, 9] build license plate detectors based on the region of interest (ROI) extraction. However, these methods are limited to the frontal or rear views, and not applicable to images with large viewpoint variations and low resolutions.

Low-resolution face recognition: low-resolution face recognition methods attempt to recognize the face identity from the low-resolution images. Compared to high resolution images, low resolution images lose the detailed information and cause the large performance drop. To improve the recognition accuracy, [25, 11] learn the mappings from low-resolution and high-resolution image pairs in the high-level feature space. [22] restore the high-resolution images from the low-resolution images to improve the classification network. [10] transfer the near-infrared spectrum (NIR) images to visible spectrum (VIS) images to tackle the poor quality of NIR images. Different from the above methods that reconstruct the images from the same domain, our method aims at improving the fine-grained object recognition from different domains.

Fine-grained vehicle classification and verification: existing fine-grained object recognition methods often focus on high resolution web images. [8] and [12] fit 3D models of vehicles into 2D images for better extracting the part-based features and rectifying the vehicle pose. [7] learn the discriminative parts of vehicles for fine-grained classification. [23] released the first large-scale fine-grained vehicle dataset, which contains 163 car makes and 1716 car models and with their fine-grained attributes. Similar to our work, [20] address the problem of fine-grained vehicle recognition in the surveillance environments. They released a large-scale vehicle dataset collected by surveillance cameras, which contains 2D and 3D bounding box labels. By unpacking the 3D bounding boxes into a 2D plane, vehicle parts are better aligned. They use the rasterized 3D bounding boxes and vehicle viewpoints as the additional inputs to further improve the recognition accuracy. Since 3D bounding boxes or 3D models cost extra computation resource and and may not be always available, we focus on how to utilize relatively accessible 2D resources. Our experiments demonstrate that our approach shows better performance compared to their methods while does not requiring any 3D information.

Unsupervised domain adaptation: [21] and [2] learn a discriminative mapping of target images to the source feature space in an unsupervised domain adaption setting. Source classifiers can be directly applied in the target domain. [18] propose a refiner neural network that improves the realism of the synthetic images by using the unlabeled image data. Note that different from the conventional domain adaptation settings, where the performance is often bounded by using all the supervised information in the target domain. Here, we focus on how to leverage the domain knowledge to improve the performance in target domain in a supervised manner. Our experiments show that the proposed hallucination network achieves even better perfor-

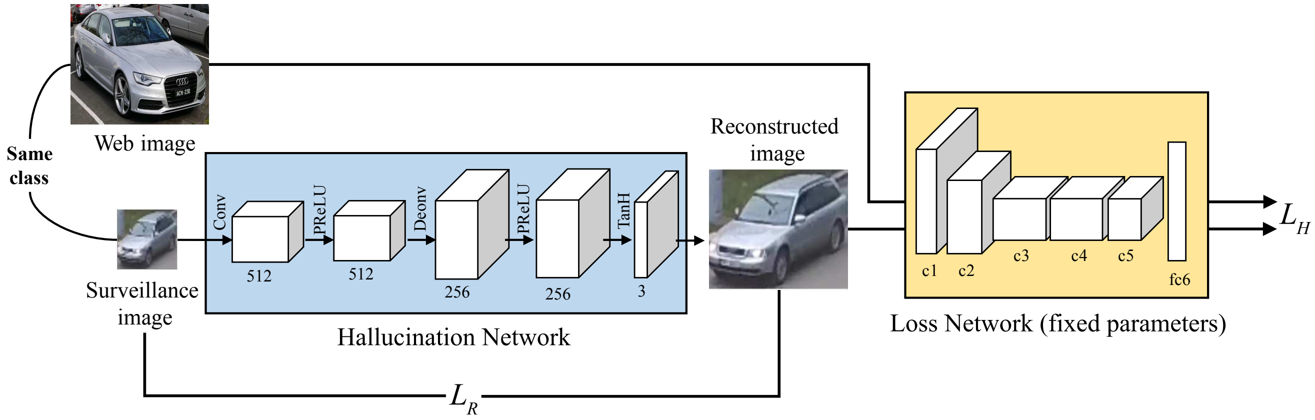

Figure 3. An overview of our hallucination network architecture. We first resize the input surveillance images by a bicubic interpolation to unify the input image size, and apply a set of non-linear transformations to generate high-resolution reconstructed images, which are two times larger than the input images. The hallucination network is trained with surveillance and web image pairs from the same class. We optimize the loss L_H to minimize the high-level CNN feature distances of the input pairs, and adopt L_R loss to preserve the color and texture information from the original image.

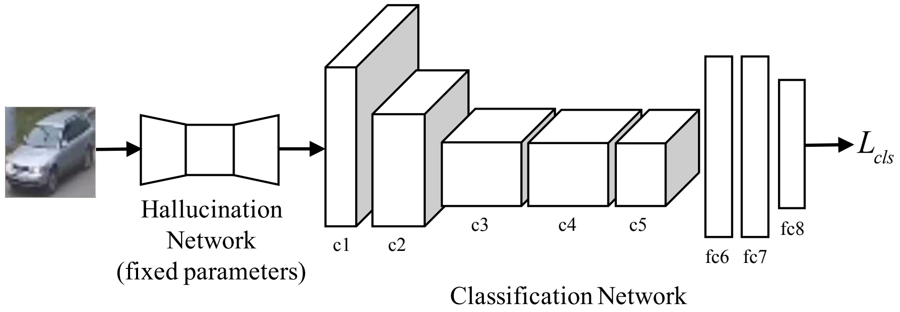

Figure 4. For training the classification network, we fix the parameters of hallucination network and optimize the softmax loss L_{cls} of the classification network.

mance compared to using all the supervised label information in target or in both source and target domains.

3. Web vehicle dataset

It is not straightforward to collect a large scale fine-grained vehicle dataset from surveillance videos, as images are low resolution and often bias to the certain classes. In contrast, it is easier to retrieve images with different models and types from Internet, e.g., crawling the images by searching the keywords. The vehicle images from Internet are often high-resolution and thus beneficial to the fine-grained object classification. For each vehicle class in the surveillance dataset, we collect the corresponding vehicle images from Internet as the auxiliary data. We use images from two different domains of the same vehicle class for training the hallucination network (see the left side of Figure 3). The hallucination network minimizes the discrepancy between web and surveillance domains and improves the quality of low resolution surveillance images.

4. Proposed method

4.1. Hallucination network

Vehicle images captured by the surveillance cameras are often low resolution and blurry. We propose a cross-domain hallucination network to better recover the detailed parts of the surveillance vehicle images. Figure 3 shows the architecture of the proposed hallucination network. It takes a surveillance image as input and generates a two-times larger reconstructed image. Different from the previous methods for face recognition [22][24], where the hallucination networks are trained on the low and high resolution image pairs from the same domain. In the surveillance environments, we do not have such corresponding high resolution images (i.e., we do not have paired low and high resolution images). To tackle this problem, we collect an additional fine-grained vehicle dataset from Internet and use them as the pseudo high resolution images for training the hallucination network. The web vehicle images often have different poses, colors and textures than the original low-resolution images, which makes our problem even more challenging.

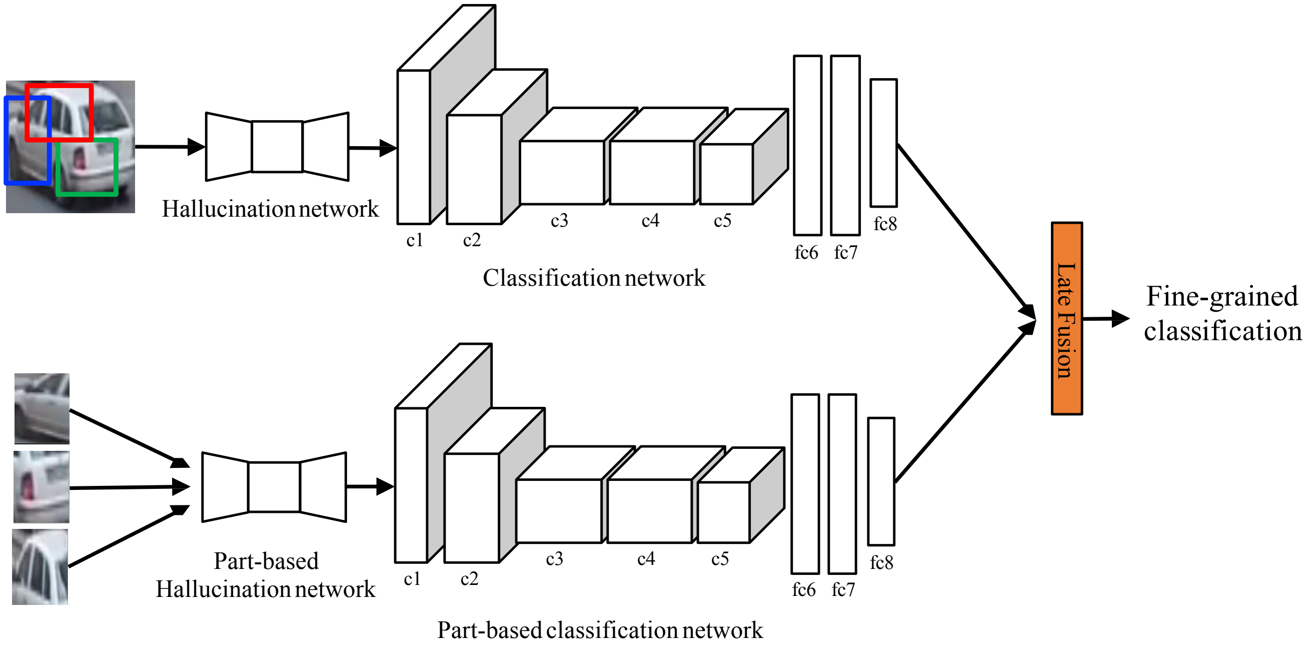
Hallucination network
c1
c2
c3
c4
c5
fc6
fc7
fc8
Classification network
Late Fusion
Fine-grained classification
Part-based Hallucination network
c1
c2
c3
c4
c5
fc6
fc7
fc8
Part-based classification network

Figure 5. We ensemble the whole and part-based hallucination and classification networks for obtaining the final classification results. The part-based hallucination network is trained by using different part images extracted from the original vehicle image, where the network structure is similar to Figure 3. All part regions share one hallucination network. We combine the results from both networks by using the weighted softmax, and obtain the predicted class with the maximum probability.

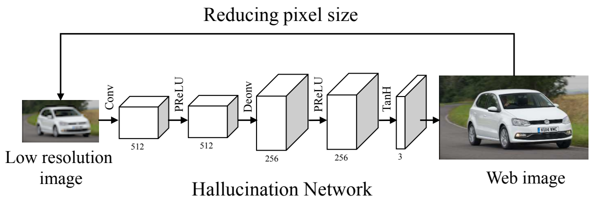

Figure 6. We pre-train the hallucination network by using the low and high resolution web image pairs, where the low resolution images are obtained by down-sampling the high-resolution web images.

The core idea of our hallucination network is to minimize the domain discrepancy between the web and surveillance image pairs, where each pair belongs to the same vehicle class. Since the web and surveillance image in each pair are visually different, it is impossible to compare them in pixel level. We adopt the idea from [6] and use the concept loss network to compare each pair in feature level. Loss network is a network proposed to extract the image features and these output features will be the inputs for the final loss function. In our paper, loss network is pre-trained on ImageNet and always fixed. We use CaffeNet (from conv1 to fc6) as the backbone model of loss network to extract the CNN features for comparison. We extract the high-level CNN features from the loss network for both the reconstructed and web images and minimize the feature differences by a hallucination loss (see the right side of Figure 3):

$$L_H = \|I_w - I_s\|_2^2 \,, \tag{1}$$

Let $\phi(x)$ be the features extracted from the loss network when processing the image x, I_w be the input web image and I_s be the input surveillance image. For hallucination network, we adopt the similar architecture as FCN [14] for keeping the spatial resolution. Hallucination network mainly consists of two stages. First, the input images are first convolved with a set of convolution layers to obtain the high-level features. Second, the extracted features are then used to reconstruct the output images by a deconvolution layer. We use the down-sampled and the original high-resolution web images as the input pairs for pre-training the hallucination network (see Figure 6).

We observe that only using the hallucination loss L_H for training does not yield good reconstruction results. We conjecture that it is because the loss network is pre-trained by ImageNet, which only retains the edge and shape but discards the texture and color information of the input image. Therefore, the output images of the hallucination network tend to be grayscale and blurry. To better preserve the texture and color information of the original images, we add an additional restoration loss L_R to regularize the output image. The restoration loss is defined as:

$$L_R = \|I_s - P(H_s)\|_2^2 \,, \tag{2}$$

where I_s is the input image, H_s is the output image from the hallucination network and the function P is a max-pooling operation with stride size 2 and padding size 0. The restoration loss can be seen as a constrain to prevent the network over-fit to the hallucination loss. Two losses are equally weighted in our experiments.

4.2. Classification network

We propose a two-stage training strategy to ensemble the hallucination network and classification network. In the first stage, the hallucination network is trained by minimizing the loss L_H and L_R while fixing the parameters of loss network. For the second stage, we fix the parameters of hallucination network and optimize the softmax loss of the classification network for fine-grained recognition (see Figure 4).

4.3. Part-based hallucination networks

Part extraction. In fine-grained classification, part-based representation is often used to better associate the objects across different viewpoints. We are interested in whether the part-based representation benefits to the cross-domain hallucination network. Moreover, precise part regions should better align the vehicle images from different domains, e.g., reducing the background noise. For extracting the part regions, we manually select the some of the channels with maximum responses from certain convolution layer (we use conv5 in our experiments). The selected channels correspond to the concepts of an vehicle [1] and help us to localize important parts from vehicles. Figure 7 shows an example of part extraction from the convolution responses. From this example, we can see that certain feature maps capture the locations and patterns of parts. The locations of the cells with the max responses in the feature map are mapped back to the input image proportionally. We crop a bounding box around each part to get the part regions.

Training the network. For training the part-based hallucination networks, we use the same architecture and training strategy as the hallucination network in Figure 3 to enhance the quality of parts. All parts share one hallucination network during the training and testing (See Figure 5). In the first stage, the part-based hallucination network is optimized by hallucination loss L_H and restoration loss L_R with different part image pairs. In the second stage, we fix the model parameters of part-based hallucination network and train the classification network by optimizing the softmax loss for fine-grained classification.

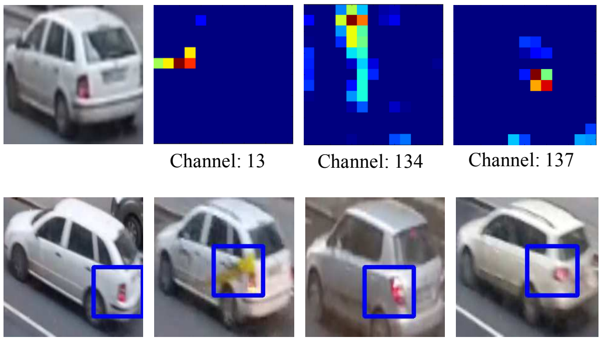

Figure 7. An example to demonstrate the part extraction from the convolution responses. **Top row:** the feature responses from the selected channels in conv5. **Bottom row:** we show part extraction results by using the channel 137 for different vehicle images, where we extract the part regions by mapping back the locations in the feature map to the original input image and draw a bounding box around each point.

Model	Accuracy
Train on target	72.0%
Train on source and target	70.2%
Siamese	75.1%
Two-stream (ours)	75.5%
3D Box [20]	76.4%
Hallucination network (ours)	**77.1%**
Late fusion (ours)	**77.9%**

Table 1. Classification comparison on BoxcCar21k in medium level. The source and target are web and surveillance domain respectively. Our method outperforms the baseline method [20] while does not requiring any 3D information, and significantly improve the results that fine-tune the network by using all the target labels or using both source and target labels.

4.4. Fusing image parts and content

To further improve the recognition accuracy, we fuse the results from the whole and part-based hallucination networks in a late fusion scheme. For each network, the normalized class probability is computed by a softmax layer after the fc8. Then, the final class probability $\mathbf{prob}$ can be easily computed by a weighted late fusion:

$$\mathbf{prob} = \lambda \mathbf{prob}_{whole} + (1 - \lambda)\mathbf{prob}_{part}, \tag{3}$$

where $\mathbf{prob}_{whole}$ is class probability of whole image framework, $\mathbf{prob}_{part}$ is the class probability of part-based framework and we linearly search the λ from 0 to 1 to obtain the fusion weights.

5. Experiments

5.1. Dataset

BoxCars21k dataset [20] contains the vehicle images from the surveillance videos annotated with fine-grained labels. The dataset is divided into the classification and verification sets. We evaluate our method on the classification set in all of our experiments. Classification set is further

[1] More accurate parts can be selected by using the existing unsupervised method [19], but we leave it to future work.

Model	Accuracy
Train on target	66.9%
Train on source and target	68.7%
Siamese	72.3%
Two-stream (ours)	71.8%
3D Box [20]	73.1%
Hallucination network (ours)	**72.8%**
Late fusion (ours)	**73.6%**

Table 2. Classification comparison on BoxcCar21k in hard level.

split into medium level (40152 training images, 19590 testing images and 77 classes) and hard level (37689 training images, 18939 testing images and 87 classes). Hard level contains more similar classes (same car make but different car model) than medium level.

5.2. Experimental settings

We collect the corresponding images from Internet as the additional data. We search the relevant web images by using the vehicle class names as keywords. Then, we apply Faster R-CNN [17] to automatically extract those images containing cars. After the preprocessing, the total number of the web images is 43,245. We use the CaffeNet pre-trained by ImageNet [4] as our backbone model, following the same settings as [20] for fair comparison. We train our method with 16 image pairs per batch, learning rate of 0.0001, iterations of 15000, momentum of 0.9, weight decay of 0.0005 and adopt Adam solver. For hallucination network, we normalize the pixel values to [-1,1] by $\hat{p} = (p - 127)/255$, where the normalized pixel value and input pixel value are denoted as $\hat{p}$ and p respectively. No other data argumentation tricks are applied in our method.

5.3. Baseline methods

We compare our approach with a set of baseline methods to verify the feasibility of the cross-domain hallucination network. In the following, we use the source and target for web and surveillance domain respectively. **(1) Train on target:** we fine-tune the CaffeNet by using the image labels from the target domain. **(2) Train on source and target:** we fine-tune the CaffeNet by using both source and target labels. We are interested to see whether using source domain labels can further improve the classification accuracy in target domain. **(3) Siamese network:** Our implementation is based on the siamese model from Caffe, which uses weight sharing and a contrastive loss function at fc6 layer. We extend the original siamese model and incorporate a softmax layer on the top of each stream for fine-grained classification. **(4) Two-stream network (ours):** we observe that there exists large domain differences between the web and surveillance domains, which implies that two domains may not share the common visual feature representations. Instead, we use two CaffeNets without the weight sharing to learn the individual visual concept for each stream. Though without the weight sharing, the high level layers are still regularized by the contrastive loss and thus the model is still able to transfer the knowledge across domains. **(5) 3D box [20]:** the method extracts the 3D bounding boxes of each vehicle by using three vanishing points. Vehicle images can be better aligned by unpacking the 3D bounding boxes into 2D planes. Using the rasterized 3D bounding boxes images (2D array) and veiwpoints (1D vector) as the additional inputs to improve the performance in hard level.

5.4. Experimental results

The experimental results are shown in Table 1 and Table 2 for medium and hard level respectively. Our method shows the better results compared to fine-tuning the CaffeNet on the target domain (train on target) and even on both source and target domains (train on source and target). The reason that using all labels from both domains even gets worse results compared to fine-tuning the model on the target only is there may exist a large difference between the source and target domain (e.g., resolution, lighting, pose, background), and an unified network can not fit both domains well.

Two-stream model shows slightly better results compared to siamese (75.5% vs 75.1%), which demonstrates the utility of using two different networks for better handling the large domain differences. Compared to [20] that estimates three vanishing points for obtaining the 3D bounding box of a vehicle, our method only requires 2D bounding boxes, which is much easier to obtain (e.g., by using current object detection methods [16, 13]). The results show that our method achieves better performance, and more importantly it is generalizable to different scene structures without the need to estimate 3D geometry.

We combine the whole image-based hallucination network and part-based hallucination network by a late fusion scheme, which shows the best performance compared to other approaches. We first combine the class probabilities of three vehicle parts by simply averaging, as three parts achieve the similar performance. Then, we combine the class probabilities of the whole image and parts by using the weighted softmax fusion. The late-fusion result shows that part-based hallucination network provide the complementary information, and further boost the final classification performance. Since using parts as inputs lose some global information (eg, shape, pose), we found that part-based framework does not yield the same or better result than the whole image framework. Thus, part-based framework works as a auxiliary role in our entire model. The improvement by late fusion scheme confirms that part-based hallucination network is helpful to the whole problem. In addition, the two testing set in our experiments has nearly 20000 vehicle images. Even 0.8% improvement (from Hallucination network to Late fusion) indicates about 160 testing examples been corrected. This shows that the improvement is robust and convincing.

In Table 3, we show an example to demonstrate that better image quality obtained by our hallucination network is crucial for fine-grained object classification. When using the baseline model CaffeNet, the vehicle class Volkswagen Polo is often wrongly classified as Skoda Citigo or Skoda Octavia Combi. Our hallucination network enhances the image details and significantly improves the accuracy by 16.5% with only slight performance drop for another class.

	VW Polo	Skoda Citigo	Skoda Octavia Combi
CaffeNet	0.598	0.892	0.930
Hallucination network	**0.763**	0.928	0.910

Table 3. An example to demonstrate that our hallucination network enhances the image quality and significantly reduces the misclassification for confusing fine-grained categories.

6. Conclusions

We propose a cross-domain hallucination network for fine-grained vehicle recognition in the surveillance environments. Our method better recovers the details and enhances the quality of the low resolution images. We also investigate the part-based hallucination network to better associate the cross-domain regions and improve the performance. In future work, we will focus on integrating the classification and hallucination networks into a joint model and train it in an end-to-end manner.

7. Acknowledgement

This work was supported in part by MediaTek Inc and the Ministry of Science and Technology, Taiwan, under Grant MOST 107-2634-F-002-007. We also benefit from the grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

[1] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas. A license plate-recognition algorithm for intelligent transportation system applications. *TITS*, 2006.

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CoRR*, abs/1612.05424, 2016.

[3] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *TITS*, 2004.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei. Fine-grained car detection for visual census estimation. In *AAAI*, 2017.

[6] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[7] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei. Learning features and parts for fine-grained recognition. In *ICPR*, 2014.

[8] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, 2013.

[9] V. Lajish and S. K. Kopparapu. Mobile phone based vehicle license plate recognition for road policing. *CoRR*, abs/1504.01476, 2015.

[10] J. Lezama, Q. Qiu, and G. Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. *arXiv preprint arXiv:1611.06638*, 2016.

[11] B. Li, H. Chang, S. Shan, and X. Chen. Low-resolution face recognition via coupled locality preserving mappings. *IEEE Signal processing letters*, 2010.

[12] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, 2014.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[15] R. Parisi, E. Di Claudio, G. Lucarelli, and G. Orlandi. Car plate recognition by neural networks and image processing. In *IEEE International Symposium on Circuits and Systems*, 1998.

[16] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[18] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *CoRR*, abs/1612.07828, 2016.

[19] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *ICCV*, 2015.

[20] J. Sochor, A. Herout, and J. Havel. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In *CVPR*, 2016.

[21] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464, 2017.

[22] J. Wu, S. Ding, W. Xu, and H. Chao. Deep joint face hallucination and recognition. *CoRR*, abs/1611.08091, 2016.

[23] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015.

[24] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016.

[25] W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *TIP*, 2012.