A cycle-consistent reciprocal network for visual correspondence

Zhiquan He[a,b,c], Donghong Zheng[b], Wenming Cao[b*]
[a] College of Information Engineering, Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China; [b] Guangdong Multimedia Information Service Engineering Technology Research Center, Shenzhen University, Shenzhen, China; [c] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, China

ABSTRACT

Visual correspondence refers to building dense correspondences between two or more images of the same category. Ideally, the predicted keypoints output by the model can be back to the source image's keypoints through the same type of network. However, in practical situations, the predicted keypoints usually do not perfectly map back to the source image keypoints. In order to strengthen the cycle-consistency of the model, we propose a cycle-consistent reciprocal network. The network uses joint loss functions to alternately train forward and inverse models, which makes the two models subject to cycle constraints and perform better with the help of each other. Experiment results demonstrate the performance of the model is improved on three popular benchmarks and set a new state-of-the-art on the benchmark of PF-WILLOW.

Keywords: Cycle-consistency, visual correspondence, reciprocal network

1. INTRODUCTION

Establishing visual correspondence as a fundamental problem has long been concerned by the computer vision community. The task aims to establish pixel-level correspondences between two or more semantically similar images, which have proven useful in a variety of applications such as object detection[1], scene understanding[2], and semantic segmentation[3]. With the development of deep networks and abundant data available, great breakthroughs have been made in representation learning for establishing visual correspondences[4, 5]. However, it is still challenging to establish visual correspondences because there are large intra-class variations between images of the same class due to illumination, scale, translation, blur and occlusion, etc.[6-9].

Geometric constraint is an effective method to reduce the number of uncertain candidate regions and is adopted by many methods. Recent approaches use neighbourhood consensus[10-12] to establish semantic correspondence. The first learnable neighbourhood consensus network (NC-Net)[10] used 4D tensor to store pixel-level matching scores and refined it through neighbourhood consensus based on local spatial context. Li et al.[11] developed NC-Net to adaptive neighbourhood consensus network (ANC-Net) with the kernels of non-isotropic 4D convolution. Jae Lee et al.[12] introduced Patch-Match Neighbourhood Consensus (PMNC), which used PatchMatch[13] to find the candidate regions with the highest similarity iteratively. However, these approaches mainly focus on translation and heavily influenced by the quality of the original correlation map under features representation.

To address the problems above, the latest methods[14-17] consider enhancing the feature representation. Cho et al.[16] pay attention to the stage of cost aggregation, aiming to reduce the effect of background clutter and achieve global consensus among refined correlation maps. Zhao et al.[14] propose a multi-scale matching network (MMNet) to enhance the network's ability of handling scale changes. Convolutional Hough Matching Networks (CHMNet)[17] extern 4D convolution to 6D convolution, adding the dimension of scale. But they ignore the cycle-consistency of the model during training, for the predicted position should be back to the starting position of the source image through the same type of network.

In this work, we introduce a cycle-consistent reciprocal network to improve the performance of CHMNet, and the improved model called CCR-CHM. Cycle consistency is a simple but useful technique in machine learning, which can also be used for training visual correspondence models. Taking a pair of cats as an example, we hope to build the map

* wmcao@szu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125064V
 · 0277-786X · doi: 10.1117/12.2662501

from the source image cat eyes to the target image cat eyes, and the predicted target position also should be mapped back to the source image cat eyes through an inverse network. In order to build this mapping relationship, we use cycle constraints to train forward and inverse networks alternately so that the two networks can be improved together during training. The experiment result shows that the model has a great improvement in three standard benchmark datasets and sets a new state-of-the-art on the dataset of PF-WILLOW[18].

This paper is organized as follows. Section 2 introduces the related work. The architecture of cycle-consistent reciprocal network is presented in Section 3. Section 4 reports the experimental results compared with other methods and offers ablation studies. Finally, we make a brief conclusion in Section 5.

2. RELATED WORK

2.1 Semantic correspondence

Early works[18, 19] mostly used hand-crafted features such as SIFT and ORB to establish a semantic correspondence, which existed the disadvantage of insufficient semantic patterns and hard to deal with background clutter, non-rigid deformation and blur. Convolutional neural network (CNN) is powerful in extracting deep feature representation[20, 21], and it becomes popular in the task of semantic correspondence soon[22-25]. They set CNN that has been pretrained on image classification as the backbone module, and then use different algorithms to build a correlation map based on the output of CNN. Rocco et al.[26] propose a CNN regressor model that estimated affine transformation parameters with deep learning method. NC-Net [10] uses 4D convolution for neighbourhood consensus task, which can effectively filter unreliable matches. PMNC and ANC-Net make progress on this basis, improving the calculation efficiency and accuracy. Jeon et al.[27] construct a multiple affine network with a pyramid structure, realizing estimation from coarse to fine. Subsequent methods[14, 15, 17, 24] mostly focus on feature representation enhancement and the method to build a correlation map efficiently.

2.2 Convolutional Hough matching

The Hough transform is a classical algorithm for object detection, which can recognize objects of specified shape in an image by voting in parameter space, and it has been proven effective in non-rigid image matching[28]. Min et al. propose a trainable convolution Hough matching layer. They also set CNN as the feature extractor, and scaled the image features, extending the 4D correlation tensor to 6D. They design a trainable convolutional Hough matching kernel combined with geometric constraints and applied it to high-dimensional 4D and 6D convolutions, which makes impressive progress in the accuracy of predicting sparse keypoints.

2.3 Cycle-consistency

Cycle-consistency is widely used in practical applications. A typical application is the area of machine translation, where the model should be roughly consistent in the process of translation and back translation. In computer vision, cycle-consistency has been applied to action prediction[29], image-to-image translation[30] and dense image alignment[31]. CycleGAN[30] used a cycle-consistency loss to learn a pixel-wise mapping relationship in the area of image-to-image translation. Inspired by their work, we further propose a cycle-consistent reciprocal network and use joint loss functions to iteratively train the visual correspondence model.

3. METHOD

Let $\mathrm{X} = x_1, x_2, \ldots x_\mathrm{N}$ be the keypoints in source image I, and visual correspondence model needs to map X to the corresponding ground-truth $\mathrm{Y} = y_1, y_2, \ldots y_\mathrm{N}$ in target image I'. As shown in Figure 1, cycle-consistency for visual correspondence considers such a relationship that if the predicted positions $\widehat{\mathrm{Y}}$ in I' is perfect, it can return to the starting points X from $\widehat{\mathrm{Y}}$ through the same type of model.

To encourage cycle-consistency of the model, we propose a reciprocal network based on cycle-consistency. Figure 2 illustrates the architecture of our network. Firstly, we need to train two networks, a forward network F_θ which predicts the target positions $\widehat{\mathrm{Y}} = F_\theta(\mathrm{X})$ and an inverse network G_φ which predicts the source positions $\widehat{\mathrm{X}} = G_\varphi(\mathrm{Y})$. The training of F_θ as usual, and the G_φ needs to exchange the position of the source images and the target images when inputting the images, so that the flow estimation output by G_φ is mapped from the target images to the source images. It should be

noted that the target position is input to the model as a known condition during training, and we offer an ablation study in the 4.2 section. We hope the following mapping relationship can be established if F_θ and G_φ are well trained:

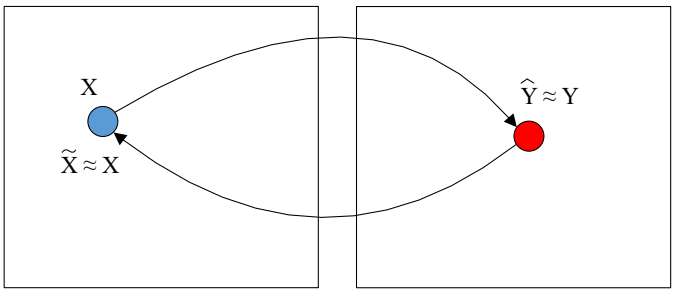

Figure 1. Illustration of cycle-consistency.

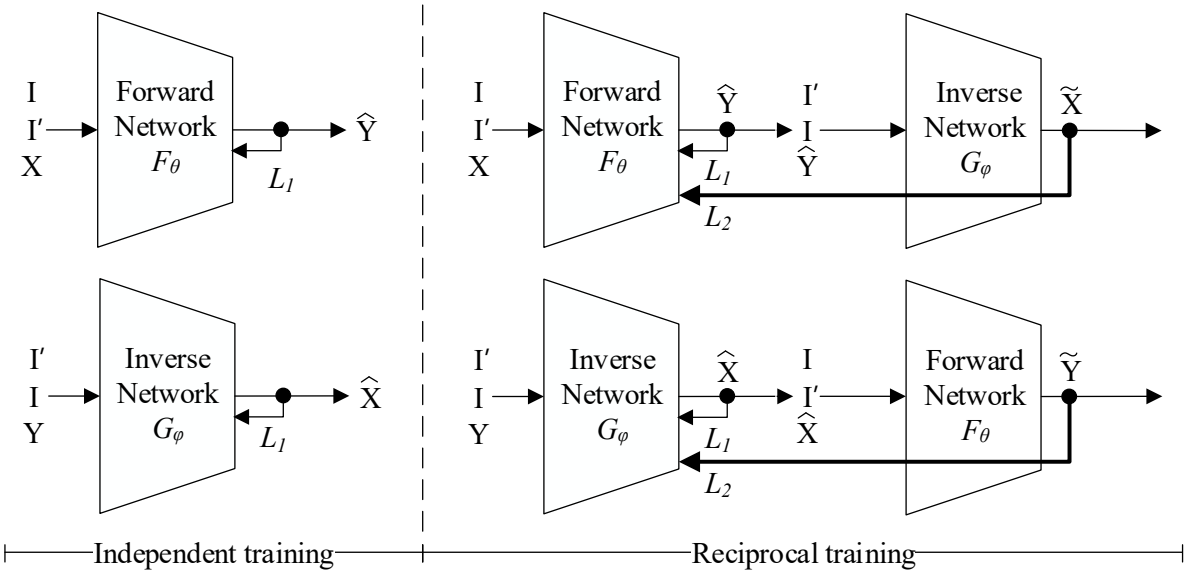

Figure 2. The architecture of cycle-consistent reciprocal network. Firstly, the forward and inverse networks are trained independently. Then, we use the first prediction loss L_1 and cycle-consistency loss L_2 to alternately train forward and inverse networks.

$$X \approx G_\varphi(F_\theta(X)) \tag{1}$$

$$Y \approx F_\theta(G_\varphi(Y)) \tag{2}$$

It is obvious that the two networks are strongly correlated and can help each to improve performance. We use cycle-consistency constraints (1) to verify the accuracy of the forward prediction $\widehat{Y}$. If the inverse network G_φ is trained well and the first prediction $\widehat{Y}$ is accurate, the second prediction $\widetilde{X}$ will be close to the starting positions X with high probability. The same explanation can be applied to equation (2). Loss function consists of the L2 loss. As shown in Figure 3, we define the loss function L_1 as the first prediction loss, and L_2 as the cycle-consistency loss, which is the loss of re-prediction based on the first prediction results.

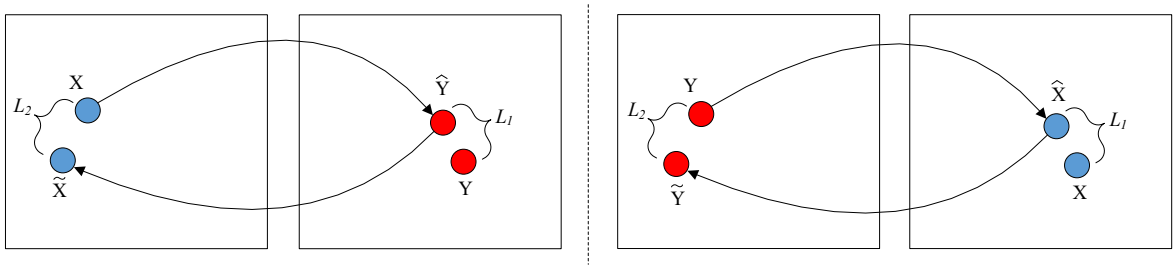

Figure 3. Illustration of joint loss.

When the forward network F_θ and inverse network G_φ are successfully trained, we come to the next step — reciprocal training. We use a reciprocal network to leverage the forward network F_θ and the inverse network G_φ, and via iterative training to make the two networks benefit from each other. We only update the parameters of one model and freezes the parameters of the opposing model during training. The training exchange epoch for F_θ and G_φ is set as 3 epochs in experiments.

To successfully train F_θ and G_φ , we define two joint loss functions J^θ and J^φ as follows. λ takes a value between 0 and 1, and we set $\lambda = 0.5$.

$$J^\theta = \lambda L_1^\theta + (1-\lambda) L_2^\theta \tag{3}$$

$$J^\varphi = \lambda L_1^\varphi + (1-\lambda) L_2^\varphi \tag{4}$$

The inverse network G_φ can be regarded as a cycle constraint to check again the accuracy of the prediction result generated by F_θ. The two networks are encouraged to be consistent through the loss function L_2, and they also should perform well under the constraint of loss function L_1. During the training of the cycle-consistent reciprocal network, the two networks are trained alternately, and their performance will have a noticeable improvement with the help of the opposing network.

4. EXPERIMENT

4.1 Datasets and metrics

PF-WILLOW[18], PF-PASCAL[32] and Spair-71k[33] are three standard benchmark datasets containing corresponding sparse annotations for visual correspondence. The most difficult dataset is SPair-71k[33] which includes 70,958 pairs of images from 18 categories with large intra-class variations. PF-WILLOW[18] and PF-PASCAL[32] respectively contain 900 pairs of images from 4 categories and 1,351 pairs of images from 20 categories. For a fair comparison, we follow the previous assessment method that trains on the training spilt of PF-PASCAL[32], and evaluate the model on the test splits of PF-PASCAL[32] and PF-WILLOW[18]. For SPair-71k[33], the model is trained on its training set and evaluated on its test set.

The percent of correct keypoints (PCK) is used as an evaluation metric. After the model outputs the predicted keypoints k_{pred}, we can get the number of correct keypoints that satisfy the condition: $\left\| k_{pred} - k_{gt} \right\|_2 \le \alpha \cdot \max(\mathrm{H}, \mathrm{W})$, where $\alpha \in \{0.05,\ 0.1\}$ denotes a threshold and k_{gt} denotes the ground-truth in target image. H and W respectively represent the height and width of the images or the bounding box of objects.

4.2 Experiment results

Figure 4 gives 3 pairs of images, which are the evaluation result on the test splits of PF-PASCAL with the threshold of $\alpha_{img} = 0.05$. The images lie in the top is the result of CHMNet, and the bottom images get from the improved network. The image on the left is the source image and the right image is the target image. The red and green lines denote wrong and correct predictions. We also mark the ground-truth in the target image with solid yellow circles.

We compare our experimental results with other methods, as shown in Table 1. The bold numbers mean the best performance, and the underlined numbers indicate the second best. Our model shows a significant improvement on each metric compared with original model, especially on the PF-WILLOW. Specifically, the PCK of $\alpha_{bbox} = 0.05$ on PF-

WILLOW increase by 3.3%, and the other one $\alpha_{bbox} = 0.1$ increase by 2.4%, which set a new state-of-the-art on PF-WILLOW. On PF-PASCAL, our model also performs better on both thresholds, and ranks second in the list.

Figure 4. Matching results at original network and improved network.

Table 1. Comparison with other methods.

Methods	SPair-71k PCK @ α_{bbox}	PF-PASCAL PCK @ α_{img}		PF-WILLOW PCK @ α_{bbox}	
	0.1	**0.05**	**0.1**	**0.05**	**0.1**
NC-Net[10]	20.1	54.3	78.9	-	-
ANC-Net[11]	-	-	86.1	-	-
HPF[23]	28.2	60.1	84.8	-	-
SCOT[34]	35.6	63.1	85.4	-	-
DHPF[15]	37.3	75.7	90.7	49.5	77.6
PMNC[12]	50.4	**82.4**	90.6	-	-
MMNet-FCN[14]	**50.4**	81.1	91.6	-	-
Cats[16]	49.9	75.4	**92.6**	50.3	79.2
CHMNet[17]	46.3	80.1	91.6	52.7	79.4
CCR-CHM (Ours)	46.7	81.1	92.3	**56.0**	**81.8**

4.3 Ablation study

We exchange the position of source and target images, which can be regarded as a method of image augmentation. Therefore, we study the effect of this behaviour on the performance of the original network. In the ablation study, we swap the position of the input images every other epoch.

In addition, we remove the reciprocal network and train the forward network by cycle-consistency only. Specifically, we replace the inverse network with the forward network, which conducts a cycle-consistency training on itself. The ablation study results are shown in Table 2, which proves that the reciprocal training network is effective.

Table 2. Ablation study of CCR-CHM.

Methods	SPair-71k PCK @ α_{bbox}	PF-PASCAL PCK @ α_{img}		PF-WILLOW PCK @ α_{bbox}	
	0.1	**0.05**	**0.1**	**0.05**	**0.1**
Baseline	46.3	80.1	91.6	52.7	79.4
+ image augmentation	46.4	80.5	91.7	53.3	79.5
+ cycle-consistency	46.2	81.0	91.3	53.1	77.0
CCR-CHM (Ours)	**46.7**	**81.1**	**92.3**	**56.0**	**81.8**

5. CONCLUSION

In this paper, we introduce a cycle-consistent reciprocal network for visual correspondence, which uses a joint loss function to train forward and inverse networks alternately. The two networks can be improved together with the help of each other. We apply our network to the training of CHMNet, and the model performs better on the test splits of the three standard benchmarks. The evaluation results show our network can be used to improve the performance of the visual correspondence model based on deep learning. And we provide ablation studies to verify our network. We believe further research on cycle-consistency can help to establish visual correspondence.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under grants 61971290, 61771322, 61871186, and the Fundamental Research Foundation of Shenzhen under Grant JCYJ20190808160815125.

REFERENCES

[1] Milletari, F., Ahmadi, S. A., Kroll, C., et al., "Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound," Computer Vision and Image Understanding, 164, 92-102(2017).
[2] Lee, J., Kim, D., Ponce, J., et al., "SFNET: Learning object-aware semantic correspondence," Proc. CVPR, 2278-2287(2019).
[3] Hur, J. and Roth, S., "Joint optical flow and temporally consistent semantic segmentation," Proc. ECCV, 163-177(2016).
[4] Hu, J., Shen, L. and Sun, G., "Squeeze-and-excitation networks," Proc. CVPR, 7132-7141(2018).
[5] Huang, G., Liu, Z., Van Der Maaten, L., et al., "Densely connected convolutional networks," Proc. CVPR, 4700-4708(2017).
[6] Seo, P. H., Lee, J., Jung, D., et al., "Attentive semantic alignment with offset-aware correlation kernels," Proc. ECCV, 349-364(2018).
[7] Kim, S., Min, D., Ham, B., et al., "FCSS: Fully convolutional self-similarity for dense semantic correspondence," Proc. CVPR, 6560-6569(2017).
[8] Ufer, N. and Ommer, B., "Deep semantic feature matching," Proc. CVPR, 5929-5938(2017).
[9] Long, J. L., Zhang, N. and Darrell, T., "Do convnets learn correspondence?," Advances in Neural Information Processing Systems, 27, (2014).
[10] Rocco, I., Cimpoi, M., Arandjelović, R., et al., "Neighbourhood consensus networks," Advances in Neural Information Processing Systems, 1651-1662(2018).
[11] Li, S., Han, K., Costain, T. W., et al., "Correspondence networks with adaptive neighbourhood consensus," Proc. CVPR, 10196-10205(2020).
[12] Lee, J. Y., DeGol, J., Fragoso, V., et al., "Patchmatch-based neighborhood consensus for semantic correspondence," Proc. CVPR, 13153-13163(2021).
[13] Barnes, C., Shechtman, E., Goldman, D. B., et al., "The generalized patchmatch correspondence algorithm," Proc. ECCV, 29-43(2010).

[14] Zhao, D., Song, Z., Ji, Z., et al., "Multi-scale matching networks for semantic correspondence," Proc. CVPR, 3354-3364(2021).
[15] Min, J., Lee, J., Ponce, J., et al., "Learning to compose hypercolumns for visual correspondence," Proc. ECCV, 346-363(2020).
[16] Cho, S., Hong, S., Jeon, S., et al., "Cats: Cost aggregation transformers for visual correspondence," Advances in Neural Information Processing Systems, 34, 9011-9023(2021).
[17] Min, J. and Cho, M., "Convolutional Hough matching networks," Proc. CVPR, 2940-2950(2021).
[18] Ham, B., Cho, M., Schmid, C., et al., "Proposal flow," Proc. CVPR, 3475-3484(2016).
[19] Yuen, C. J. and Torralba, A., "Sift flow: Dense correspondence across scenes and its applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5), 978-994(2010).
[20] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition," Proc. CVPR, 770-778(2016).
[21] Huang, G., Liu, Z., Van Der Maaten, L., et al., "Densely connected convolutional networks," Proc. CVPR, 4700-4708(2017).
[22] Han, K., Rezende, R. S., Ham, B., et al., "SCNET: Learning semantic correspondence," Proc. ICCV, 1831-1840(2017).
[23] Min, J., Lee, J., Ponce, J., et al., "Hyperpixel flow: Semantic correspondence with multi-layer neural features," Proc. ICCV, 3395-3404(2019).
[24] Seo, P. H., Lee, J., Jung, D., et al., "Attentive semantic alignment with offset-aware correlation kernels," Proc. ECCV, 349-364(2018).
[25] Kim, S., Lin, S., Jeon, S. R., et al., "Recurrent transformer networks for semantic correspondence," Advances in Neural Information Processing Systems, 31, (2018).
[26] Rocco, I., Arandjelovic, R. and Sivic J., "Convolutional neural network architecture for geometric matching," Proc. CVPR, 6148-6157(2017).
[27] Jeon, S., Kim, S., Min, D., et al., "PARN: Pyramidal affine regression networks for dense semantic correspondence," Proc. ECCV, 351-366(2018).
[28] Cho, M., Kwak, S., Schmid, C., et al., "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," Proc. CVPR, 1201-1210(2015).
[29] Pang, G., Wang, X., Hu, J., et al., "DBDNet: Learning bi-directional dynamics for early action prediction," IJCAI, 897-903(2019).
[30] Zhu, J. Y., Park, T., Isola, P., et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proc. ICCV, 2223-2232(2017).
[31] Zhou, T., Lee, Y., Yu, S. X., et al., 2015 "Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences," Proc. CVPR, 1191-1200(2017).
[32] Ham, B., Cho, M., Schmid, C., et al., "Proposal flow: Semantic correspondences from object proposals," IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(7), 1711-1725(2017).
[33] Min, J., Lee J., Ponce, J., et al., "Spair-71k: A large-scale benchmark for semantic correspondence," arXiv preprint arXiv:1908.10543, (2019).
[34] Liu, Y., Zhu, L., Yamada, M., et al., "Semantic correspondence as an optimal transport problem," Proc. CVPR, 4463-4472(2020).