Semantic-Aware Fine-Grained Correspondence (Supplementary Material)

Yingdong Hu[1], Renhao Wang[1], Kaifeng Zhang[1], and Yang Gao[1,2]*

[1] Tsinghua University
[2] Shanghai Qi Zhi Institute
{huyd21,wangrh21,zhangkf19}@mails.tsinghua.edu.cn
{gaoyangiiis}@tsinghua.edu.cn

A Implementation Details

A.1 Fine-Grained Correspondence Network Pre-training

The implementation details of our fine-grained correspondence network are as follows.

Data Augmentation We use only spatial augmentation, where two random crops with scale $[0.0, 1.0]$ from the image are generated and resized into 256×256.

Architectures Following [4,5,9], we adopt ResNet-18 as the backbone $\mathbf{f}$ and reduce the stride of last two residual blocks (`res3` and `res4`) to 1. The modified backbone produces a feature map with size 32×32 (ablation in Appendix B.2). The dense projection and prediction head use the same architecture: a 1×1 convolution layer with 2048 output channels followed by batch normalization and a ReLU activation, and a final 1×1 convolution layer with output dimension 256. The positive radius r used to control the size of spatial neighborhood is set to 0.5.

Optimization We train the model with the Adam optimizer for 60k iterations. The learning rate is set to 0.001. The weight decay is set to 0. The batch size is 96. For the target network, the exponential moving average parameter τ starts from 0.99 and gradually increases to 1 under a cosine schedule, following [2]. The whole model can be trained on a single 24GB NVIDIA 3090 GPU.

A.2 Label Propagation

We follow the same label propagation algorithm in [4]. Specifically, given the ground-truth labels in the first frame, a recurrent inference strategy is applied to propagate the labels to the rest of the frames: we calculate the similarity between the current frame with the first frame (to provide ground truth labels) as well as the preceding m frames (to provide predicted labels). We reduce the stride of the penultimate residual block (`res4`) of the backbone network to be 1 and use its output (stride 8) to compute a dense similarity matrix. To avoid ambiguous matches, we define a localized spatial neighborhood by computing

---

[*] Corresponding author.

the similarity between pixels that are at most r' pixels away from each other. Finally, the labels of the top-k most similarly local feature vectors are selected and are propagated to the current frame.

For a single network which only learns semantic correspondence or fine-grained correspondence, the detailed test hyper-parameters for the three datasets are listed in Table A.1.

Table A.1. Test hyper-parameters for a single network.

	DAVIS	JHMDB	VIP
top-k	10	10	10
preceding frame m	20	8	8
propagation radius r'	12	3	15

Recall that in fusing the two different kinds of correspondence, we introduce a new hyper-parameter λ. We report the test hyper-parameters for combined correspondence in Table A.2. In general, we find that more neighbors (larger top-k and propagation radius r') are required for consistent performance.

Table A.2. Test hyper-parameters when fusing two kinds of correspondence.

	DAVIS	JHMDB	VIP
weight λ	1.75	1.0	1.0
top-k	15	20	10
preceding frame m	20	8	8
propagation radius r'	15	5	15

A.3 Semantic Segmentation Protocol

The backbone is kept fixed and we train a 1×1 convolutional layer on top to predict a semantic segmentation map. We apply dilated convolutions in the last residual block to obtain dense predictions. We use PASCAL[1] `train_aug` and `val` splits during training and evaluation, respectively. We adopt mIoU as the metric. The 1×1 convolutional layer training uses base $lr = 0.1$ for 60 epochs, weight decay = 0.0001, momentum = 0.9, and batch size = 16 with an SGD optimizer.

A.4 Linear Classification Protocol

Given the pre-trained network, we train a supervised linear classifier on top of the frozen features, which are obtained from ResNet's global average pooling

Table B.3. Results on DAVIS-2017 of FC using different training datasets. The number of images per dataset is in parentheses.

Dataset	PASCAL(17K)	COCO(118K)	YT-VOS(95K)	ImageNet(1.28M)
$\mathcal{J}\&\mathcal{F}_\mathrm{m}$	67.9	68.2	67.7	67.9

layer. We train this classifier on the ImageNet train set and report top-1 classification accuracy on the ImageNet validation set. Following prior work[3], the linear classifier training uses base $lr = 30.0$ for 100 epochs, weight decay $= 0$, momentum $= 0.9$, and batch size$= 256$ with a SGD optimizer.

A.5 Combined with Image-level Pretext Task

We add BYOL loss to FC for joint optimization. Specifically, the two loss functions share the same backbone encoder (outputs a feature map with size 32×32) and data loader (performs only spatial augmentation). But the projection head and prediction head are not shared. The projection head of BYOL is a two-layer MLP whose hidden and output dimensions are 2048 and 256. Note that BYOL average-pool backbone features to aggregate information from all spatial locations. Other implementation details follow FC. Two loss functions are balanced by a multiplicative factor α (set to 1 by default).

B Additional Experimental Results

B.1 FC is Robust to Different Dataset

When pretrained on non object-centric dataset (e.g. COCO [6]), the performance of typical image-level self-supervised methods drop significantly [7,8]. At the same time, it is largely recognized that a larger dataset usually results in stronger semantic representation for these methods. But this may not be true for a task that requires analyzing low-level cues. The following Table B.3 compares different training datasets of FC. We can see that FC is robust to the size and nature of the dataset. FC can effectively learn from a relatively small dataset. It actually gains more benefits from datasets that contain more complex scenes with several objects. The results on COCO even surpass Youtube-VOS used in the main body of the paper.

B.2 Feature Resolution

We report the results of our fine-grained correspondence network (FC) using different feature resolutions in Table B.4. The performance on DAVIS improves as the resolution increases. This is intuitive, because higher resolution indicates the local feature vectors correspond to a smaller region on the original image (small receptive fields), which benefits fine-grained low-level correspondence learning. But high-level semantics require larger receptive fields to encode more holistic information.

Table B.4. Effect of feature map resolution. The results increase as resolution gets higher. We use 32 × 32 by default.

Feature Resolution	$\mathcal{J}\&\mathcal{F}_\text{m}$	$\mathcal{J}_\text{m}$	$\mathcal{F}_\text{m}$
8 × 8	63.3	61.8	64.8
16 × 16	65.2	63.4	67.0
32 × 32	67.6	64.7	70.5

B.3 Semantic Segmentation and Linear Classification

The quantitative comparison on different downstream tasks is shown in Table B.5. For a fair comparison, we use ResNet-18 as MoCo backbone. CRW surpasses MoCo on DAVIS, but is dramatically outperformed by MoCo on semantic segmentation and image classification. Note that our FC model exhibits similar properties as CRW: the learned representation is suitable for fine-grained correspondence task, but lacks high-level semantic information. When we add crucial missing semantic information, our SFC achieves significant improvements on label propagation, semantic segmentation and image classification.

Table B.5. Comparison on label propagation, semantic segmentation and linear classification.

	DAVIS			PASCAL	ImageNet
Method	$\mathcal{J}\&\mathcal{F}_\text{m}$	$\mathcal{J}_\text{m}$	$\mathcal{F}_\text{m}$	mIoU	Acc@1
MoCo	62.1	60.3	63.8	25.5	48.7
CRW	67.6	64.8	70.2	13.0	12.6
FC	67.7	64.7	70.5	15.9	16.3
SFC	69.5	66.7	72.4	30.1	51.2

B.4 Semantic Correspondence Backbone

We use MoCo as the default semantic correspondence backbone in the main experiment, but our framework is extensible to any arbitrary backbone that is capable of producing spatial feature maps. In Table B.6, we show that we can flexibly swap out the semantic correspondence backbone for any off-the-shelf self-supervised network and maintain strong performance on DAVIS. Some methods such as SimCLR and BYOL even surpass MoCo. This strongly supports our hypothesis that image-level representations in general contain information about semantic correspondences.

B.5 Fine-Grained Correspondence Backbone

In Table B.7, we replace our own FC network in SFC with another fine-grained correspondence network CRW. We find the performance generally underperforms

Table B.6. Results after replacing MoCo with alternate image-level self-supervised representation learning methods.

Combination	$\mathcal{J}\&\mathcal{F}_\mathrm{m}$	$\mathcal{J}_\mathrm{m}$	$\mathcal{F}_\mathrm{m}$
InstDis + FC	70.1	67.1	73.1
MoCo + FC	71.2	68.3	74.0
SimCLR + FC	71.4	68.6	74.2
BYOL + FC	71.3	68.4	74.1
SimSiam + FC	70.2	67.3	73.1
VINCE + FC	70.4	67.7	73.2
VFS + FC	70.7	67.8	73.7

SFC. FC is better than CRW on all evaluation metrics, as shown in Table B.5. FC is also much simpler and computationally efficient. It takes less than a day using a single GPU, but CRW reports seven days of training.

The results in Table B.7 surpass image-level self-supervised methods or CRW alone, demonstrating the benefits of considering two orthogonal correspondences and the flexibility of our framework. It enables us to explore more effective and efficient self-supervised learning methods for semantic or fine-grained representations separately.

Table B.7. Replace FC with another fine-grained correspondence model CRW.

Combination	$\mathcal{J}\&\mathcal{F}_\mathrm{m}$	$\mathcal{J}_\mathrm{m}$	$\mathcal{F}_\mathrm{m}$
InstDis + CRW	69.6	66.6	72.6
MoCo + CRW	70.6	67.8	73.4
SimCLR + CRW	70.7	68.0	73.4
BYOL + CRW	70.9	68.1	73.6
SimSiam + CRW	69.7	66.8	72.6
VINCE + CRW	70.3	67.6	73.1
VFS + CRW	70.6	67.7	73.5

C Visualization

We provide a more detailed visualization of our SFC model on several downstream label propagation tasks. In Figure C.1, we show a comparison between SFC and CRW on the visual object segmentation benchmark DAVIS-2017. Our SFC model can generally output more accurate segmentation boundaries and reduce the amount of mistakes and failures made by CRW. In Figure C.2 and Figure C.3, we provide visualizations on the human pose tracking benchmark JHMDB and the human part tracking benchmark VIP. Note that in all our experiments, no prior knowledge on human structure or object class is used. The label propagation process is solely based on feature matching.

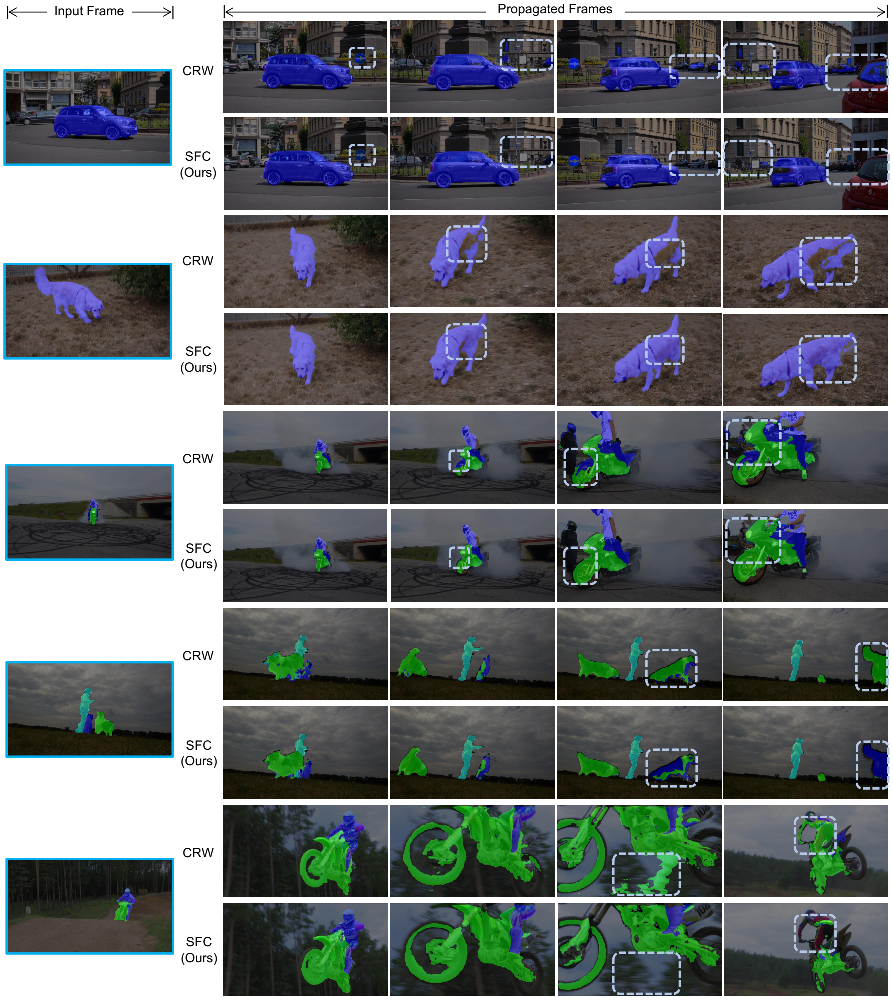

Fig. C.1. Comparing our SFC with CRW on DAVIS-2017. Within each example, the upper row is the output of CRW, and the lower row is the output of SFC. Blue dashed boxes indicate the main areas of difference.

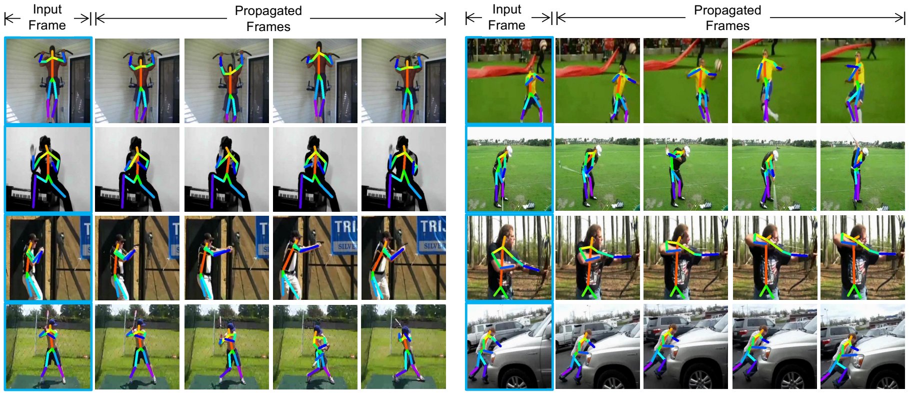

Fig. C.2. Visualization on JHMDB. Pose keypoints and their initial positions are defined on the input frame (outlined in blue), and propagated to the rest of frames.

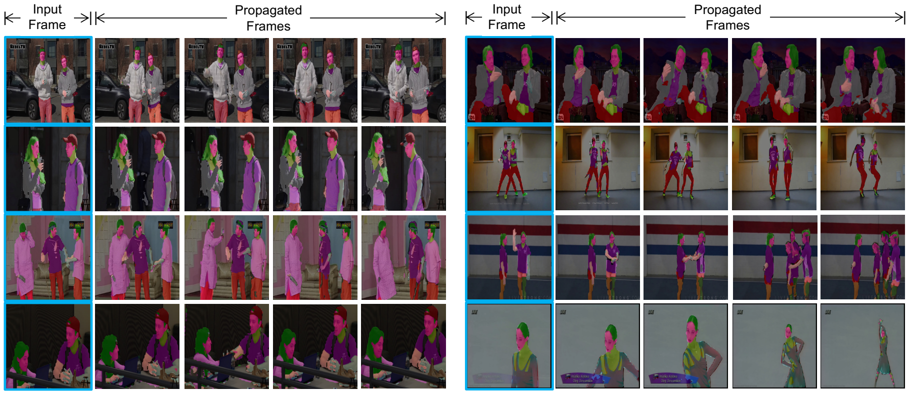

Fig. C.3. Visualization on VIP. The segmentation map of different body parts are defined on the input frame (outlined in blue), and propagated to the rest of frames.

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
2. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
3. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
4. Jabri, A., Owens, A., Efros, A.A.: Space-time correspondence as a contrastive random walk. Advances in Neural Information Processing Systems (2020)
5. Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.H.: Joint-task self-supervised learning for temporal correspondence. arXiv preprint arXiv:1909.11895 (2019)
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
7. Purushwalkam, S., Gupta, A.: Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. arXiv preprint arXiv:2007.13916 (2020)
8. Selvaraju, R.R., Desai, K., Johnson, J., Naik, N.: Casting your model: Learning to localize improves self-supervised representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11058–11067 (2021)
9. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)