Weekly Summary 2022-8-2 to 2022-8-9

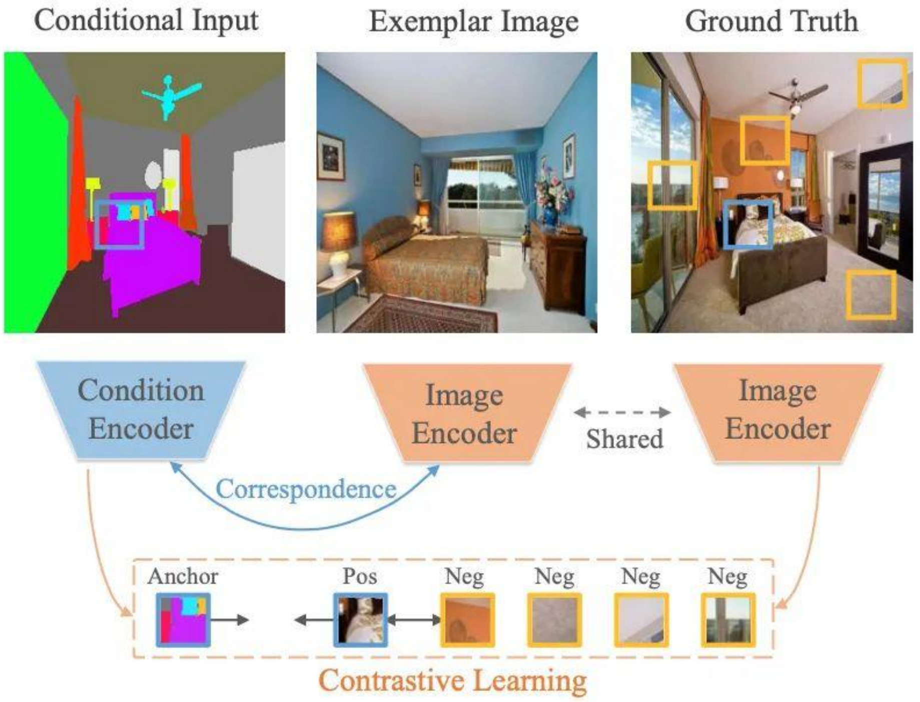
Abstract

1. Read 5 papers from CVPR2022 oral.
2. Test Relation in VOC2007 dataset.

Section 1 Paper Reading

【1】Marginal Contrastive Correspondence for Guided Image Generation

paper：https://arxiv.org/abs/2204.00442

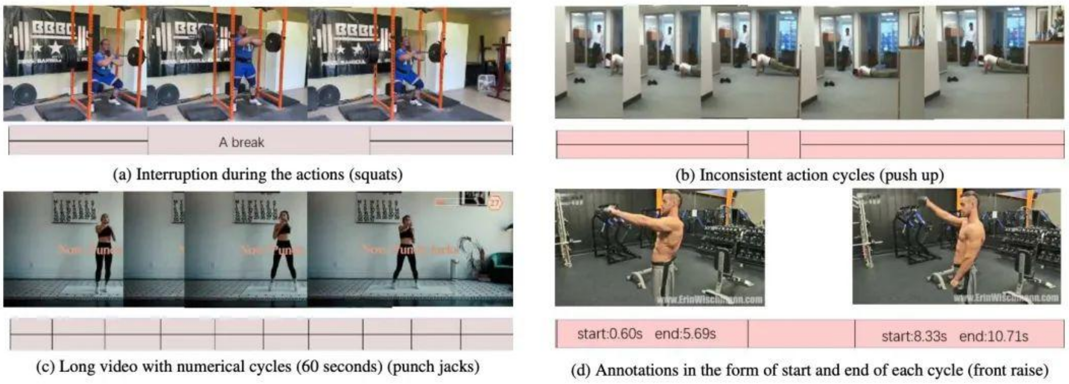


Example based image translation establishes a dense correspondence between conditional inputs and examples (from two different domains) to leverage detailed example styles for realistic image translation. Existing works implicitly establish cross-domain correspondence by minimizing the feature distance between two domains. Without explicitly exploiting domain invariant features, this approach may not be effective in reducing domain gaps, which often leads to suboptimal correspondence and image translation.

In this paper, a Marginal contrast learning network (McL-net) is designed to learn domain-invariant features through contrast learning for image translation based on real examples. Specifically, the authors devise a marginal contrast loss of innovation that guides the explicit establishment of dense correspondences. However, only establishing correspondence with domain-invariant semantics may damage texture patterns and lead to degraded texture generation quality. Therefore, the authors designed an autocorrelation graph (SCM), which incorporates the scene structure as auxiliary information and greatly improves the constructed correspondence. Quantitative and qualitative experiments on various image translation tasks show that the proposed method consistently outperforms state of the art methods.

【2】TransRAC: Encoding Multi-scale Temporal Correlation with Transformers for Repetitive Action Counting

paper：https://arxiv.org/abs/2204.01018
dataset：https://svip-lab.github.io/dataset/RepCount_dataset.html
code：https://github.com/SvipRepetitionCounting/TransRAC

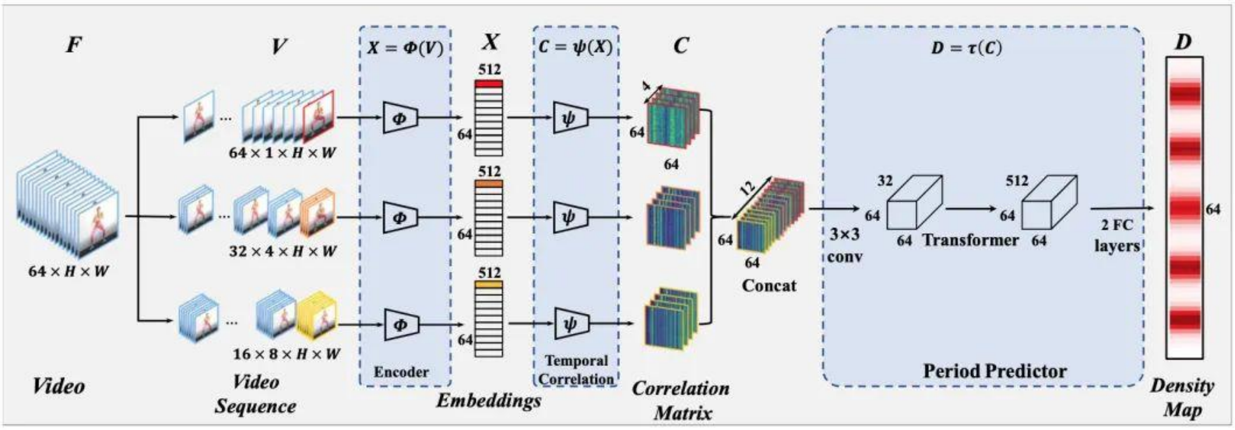


(a) Interruption during the actions (squats)

(b) Inconsistent action cycles (push up)

(c) Long video with numerical cycles (60 seconds) (punch jacks)

(d) Annotations in the form of start and end of each cycle (front raise)

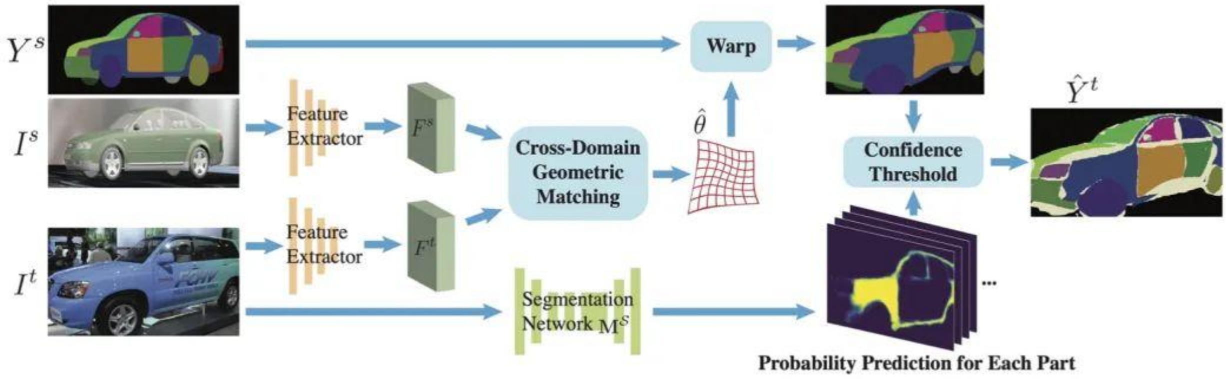


Counting repetitions is common in human activities such as physical exercise. Existing methods focus on performing repetitive action counting in short videos, which is difficult to handle in longer videos in real scenes. In the era of data flooding, this degradation of generalization ability is mainly attributed to the lack of long video datasets.

Therefore, a new large-scale repetitive action count dataset is constructed in this paper, covering various video lengths, as well as more realistic situations such as interrupted or inconsistent actions in videos. In addition, the authors provide fine-grained labels for action cycles, rather than just calculating comments and values. The dataset contains 1,451 videos and about 20,000 annotations. For repetitive actions in more realistic scenarios, the authors suggest using Transformer to encode multi-scale temporal correlations with both performance and efficiency in mind. In addition, with the help of fine-grained annotation of action cycles, this paper proposes a method based on density map regression to predict action cycles, resulting in better performance and adequate interpretability.

【3】Learning Part Segmentation through Unsupervised Domain Adaptation from Synthetic Vehicles

paper：https://arxiv.org/abs/2103.14098
dataset：https://qliu24.github.io/udapart



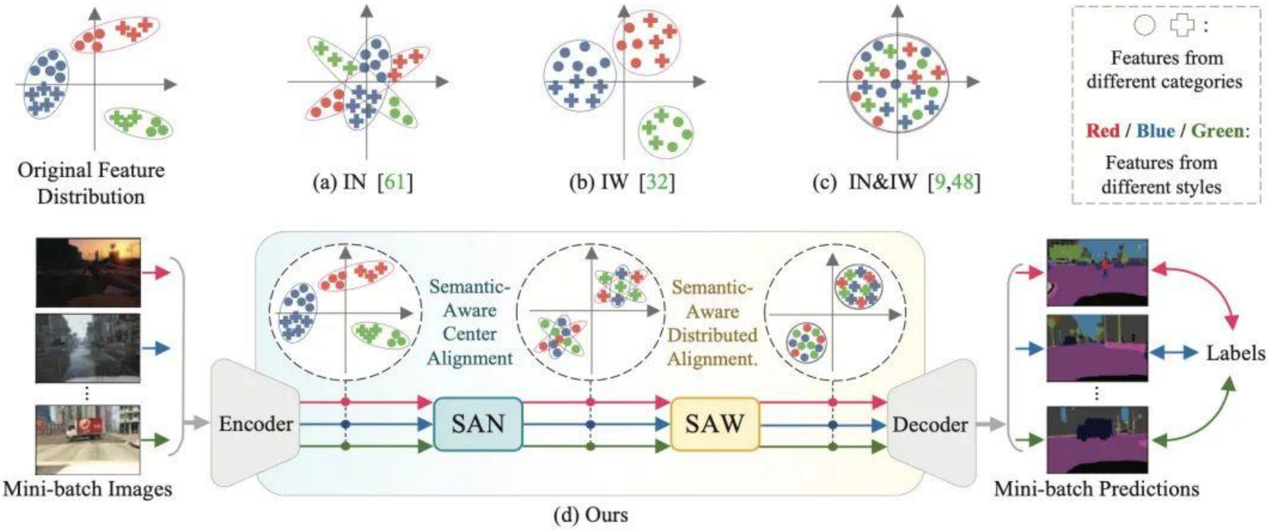
Local segmentation provides a rich and detailed description of objects at the local level. However, annotation of local segmentation requires a lot of work, which makes it difficult to use standard deep learning methods. In this paper, the authors propose the idea of learning local segmentation by unsupervised domain adaptation (UDA) in synthetic data. This paper first introduces UDA-PART, a comprehensive vehicle local segmentation dataset that can be used as a benchmark for UDA1. In UDA-Part, the author annotates parts on the 3D CAD model to generate a large number of annotated composite images. This article also annotates parts on many real images to provide a real test set. Second, to advance the adaptation of local models trained from synthetic data to real images, the authors introduce a novel UDA algorithm that uses the spatial structure of objects to guide the adaptation process. Experimental results on two real test datasets in this paper confirm that our approach outperforms existing work and demonstrate the promise of learning local segmentation of general objects from synthetic data.

【4】Semantic-Aware Domain Generalized Segmentation

paper：https://arxiv.org/abs/2204.00822
code：https://github.com/leolyj/SAN-SAW

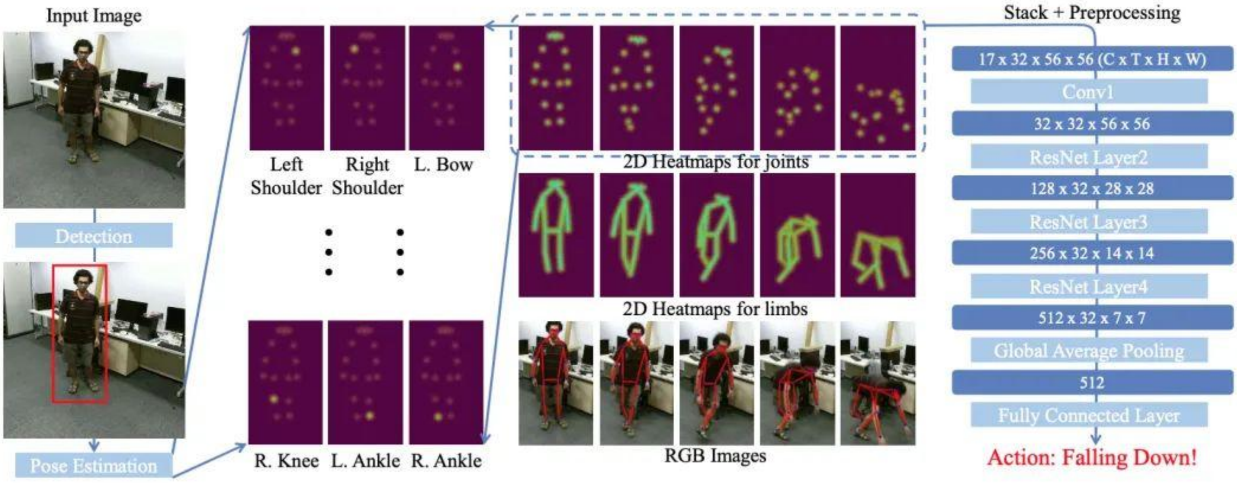
Deep models trained on source domains lack generalization when evaluated on unseen target domains with different data distributions. The problem becomes more acute when we cannot access the target domain samples for adaptation. In this paper, the authors solve the problem of domain generalization semantic segmentation, where the segmentation model is trained to be domain invariant without using any target domain data. Existing approaches to this problem normalize the data to a uniform distribution. The authors argue that although such standardization promotes global standardization, the resulting features are not discriminative enough to obtain clear segmentation boundaries.

To enhance the separation between categories while promoting domain invariance, this paper proposes a framework consisting of two new modules: semantic-aware Normalization (SAN) and semantic-aware Whitening (SAW). Specifically, SAN focuses on category-level center alignment between features from different image styles, while SAW enforces distributed alignment on features that are already center aligned. Facilitate compactness within categories and separability between categories with the help of SAN and SAW.


Original Feature Distribution
(a) IN [61]
(b) IW [32]
(c) IN&IW [9,48]
○ ✚ : Features from different categories
Red / Blue / Green: Features from different styles
Mini-batch Images
Encoder
Semantic-Aware Center Alignment
SAN
Semantic-Aware Distributed Alignment.
SAW
Decoder
Labels
Mini-batch Predictions
(d) Ours

【5】Revisiting Skeleton-based Action Recognition

paper：https://arxiv.org/abs/2104.13586
code：https://github.com/kennymckormick/pyskl



As an important feature of human movement, human skeleton has attracted more and more attention in recent years. Many bone-based action recognition methods use GCN to extract features from human bones. Although these attempts yielded positive results, the GCN-based approach was limited in terms of robustness, interoperability, and extensibility.

This work presents PoseConv3D, a novel approach for skeleton-based action recognition. PoseConv3D relies on 3D heatmap volumes rather than graphical sequences as the basic representation of the human skeleton. Compared to GCN-based methods, PoseConv3D is more effective in learning spatiotemporal features, is more robust to pose estimation noise, and generalizes better across datasets. In addition, PoseConv3D can handle multiplayer scenarios without additional computational cost. Layered features can be easily integrated with other patterns in the early fusion phase, providing significant design scope for improved performance. PoseConv3D is state of the art on five of the six standard skeleton-based action recognition benchmarks. Once fused with other modes, it achieves state of the art on all eight multimodal action recognition benchmarks.