An Overview of Self-supervised Methods

Qinwei Xu 2019/05/07

Reference:
Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey arixiv 2019

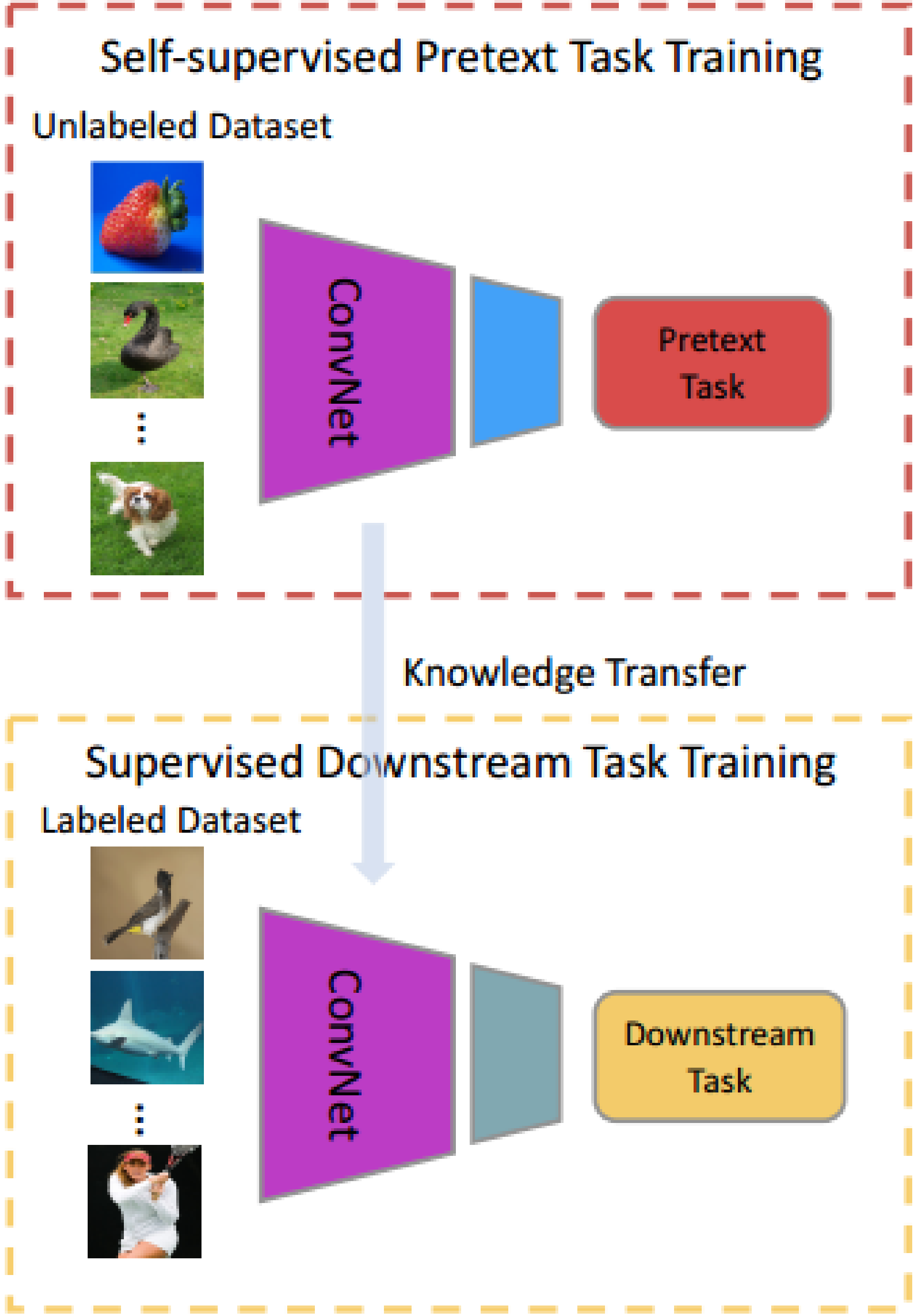
What is self-supervised learning?

- Difference between pure **unsupervised learning** and **self-supervised learning**
  - Pure unsupervised learning does not need any supervision signals
  - Self-supervised learning needs supervision signals (using automatically generated pseudo labels)
- Generalized definition
  - As long as the supervision signals are not generated by human annotations, the learning paradigm is self-supervised
- Significance
  - Target tasks can greatly benefit from self-supervised pre-training when training data (especially labelled training data) are scarce

What is self-supervised learning?

➢ General pipeline of self-supervised learning

- The purpose is to transfer features trained from self-supervised pretext tasks to supervised downstream / target tasks;
- Visual features from only the first several layers are transferred as high-level features contains task-specific signals;
- Performance of the target task is used to evaluate the quality of self-supervised learning



Self-supervised pretext tasks

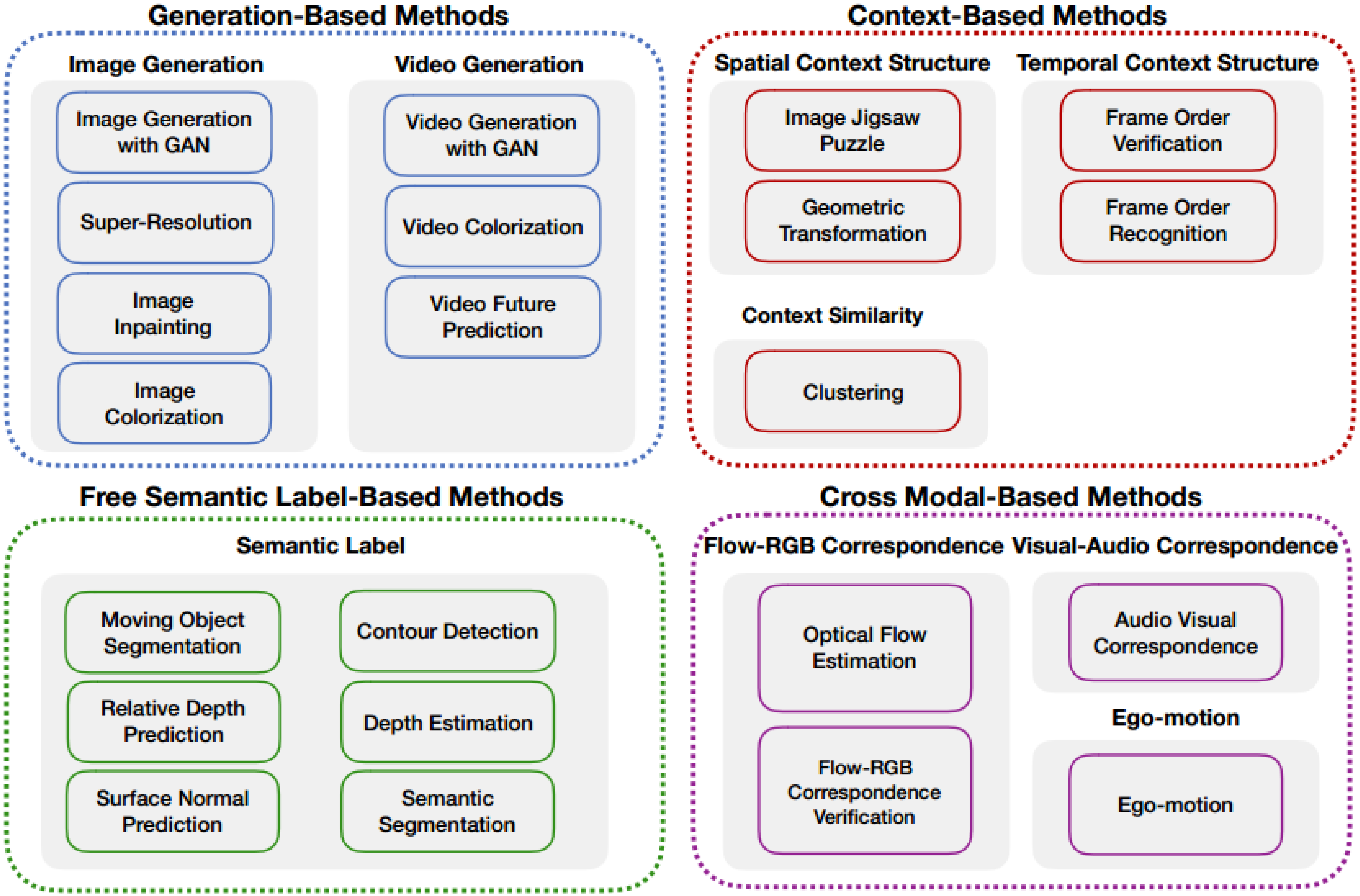


Image feature learning methods

Method	Category	Code	Contribution
GAN [83]	Generation	✓	Forerunner of GAN
DCGAN [120]	Generation	✓	Deep convolutional GAN for image generation
WGAN [121]	Generation	✓	Proposed WGAN which makes the training of GAN more stable
BiGAN [122]	Generation	✓	Bidirectional GAN to project data into latent space
SelfGAN [123]	Multiple	✗	Use rotation recognition and GAN for self-supervised learning
ColorfulColorization [18]	Generation	✓	Posing image colorization as a classification task
Colorization [82]	Generation	✓	Using image colorization as the pretext task
AutoColor [124]	Generation	✓	Training ConvNet to predict per-pixel color histograms
Split-Brain [42]	Generation	✓	Using split-brain auto-encoder as the pretext task
Context Encoder [19]	Generation	✓	Employing ConvNet to solve image inpainting
CompletNet [125]	Generation	✓	Employing two discriminators to guarantee local and global consistent
SRGAN [15]	Generation	✓	Employing GAN for single image super-resolution
SpotArtifacts [126]	Generation	✓	Learning by recognizing synthetic artifacts in images
ImproveContext [33]	Context	✗	Techniques to improve context based self-supervised learning methods
Context Prediction [41]	Context	✓	Learning by predicting the relative position of two patches from an image
Jigsaw [20]	Context	✓	Image patch Jigsaw puzzle as the pretext task for self-supervised learning
Damaged Jigsaw [89]	Multiple	✗	Learning by solving jigsaw puzzle, inpainting, and colorization together
Arbitrary Jigsaw [88]	Context	✗	Learning with jigsaw puzzles with arbitrary grid size and dimension
DeepPermNet [127]	Context	✓	A new method to solve image patch jigsaw puzzle
RotNet [36]	Context	✓	Learning by recognizing rotations of images
Boosting [34]	Multiple	✗	Using clustering to boost the self-supervised learning methods
JointCluster [128]	Context	✓	Jointly learning of deep representations and image clusters
DeepCluster [44]	Context	✓	Using clustering as the pretext
ClusterEmbegging [129]	Context	✓	Deep embedded clustering for self-supervised learning
GraphConstraint [43]	Context	✓	Learning with image pairs mined with Fisher Vector
Ranking [38]	Context	✓	Learning by ranking video frames with a triplet loss
PredictNoise [46]	Context	✓	Learning by mapping images to a uniform distribution over a manifold
MultiTask [32]	Multiple	✓	Using multiple pretext tasks for self-supervised feature learning
Learning2Count [130]	Context	✓	Learning by counting visual primitive
Watching Move [81]	Free Semantic Label	✓	Learning by grouping pixels of moving objects in videos
Edge Detection [81]	Free Semantic Label	✓	Learning by detecting edges
Cross Domain [81]	Free Semantic Label	✓	Utilizing synthetic data and its labels rendered by game engines

Generation-based image feature learning

- Pretext tasks
  - Image generation with GAN
  - Image generation with inpainting
  - Image generation with super resolution

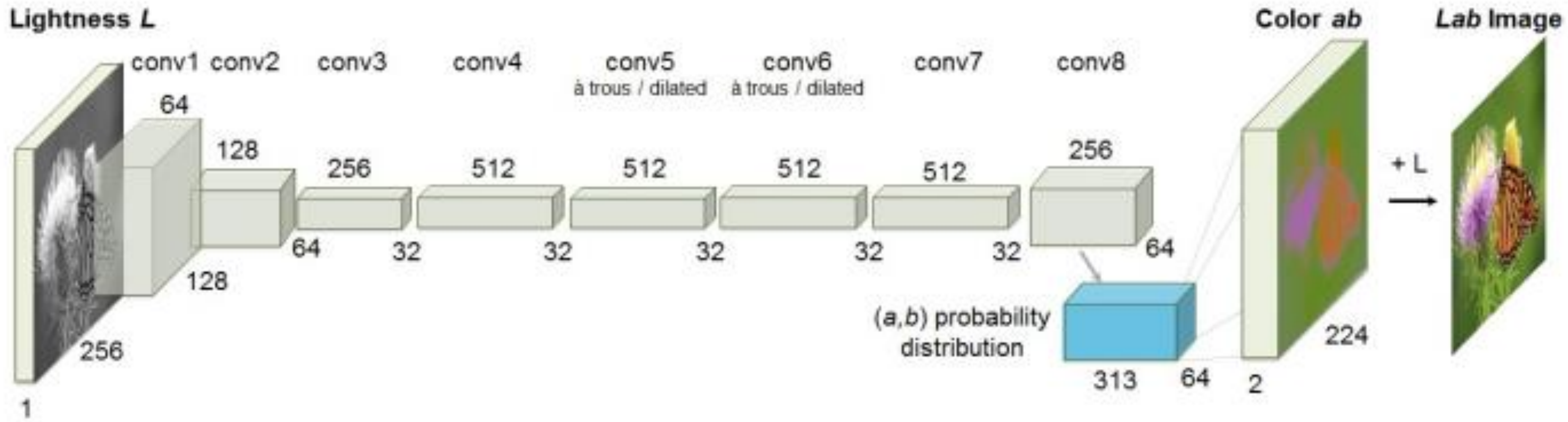
- Parameters of the discriminator can be transferred
  - It needs to capture semantic features to distinguish real or fake images

- These tasks are not designed for self-supervised pre-training
  - The main purpose is image generation / inpainting / super resolution

- Only a few works have been done to transfer the features of these tasks

Generation-based image feature learning

- Pretext tasks
  - Image generation with colorization



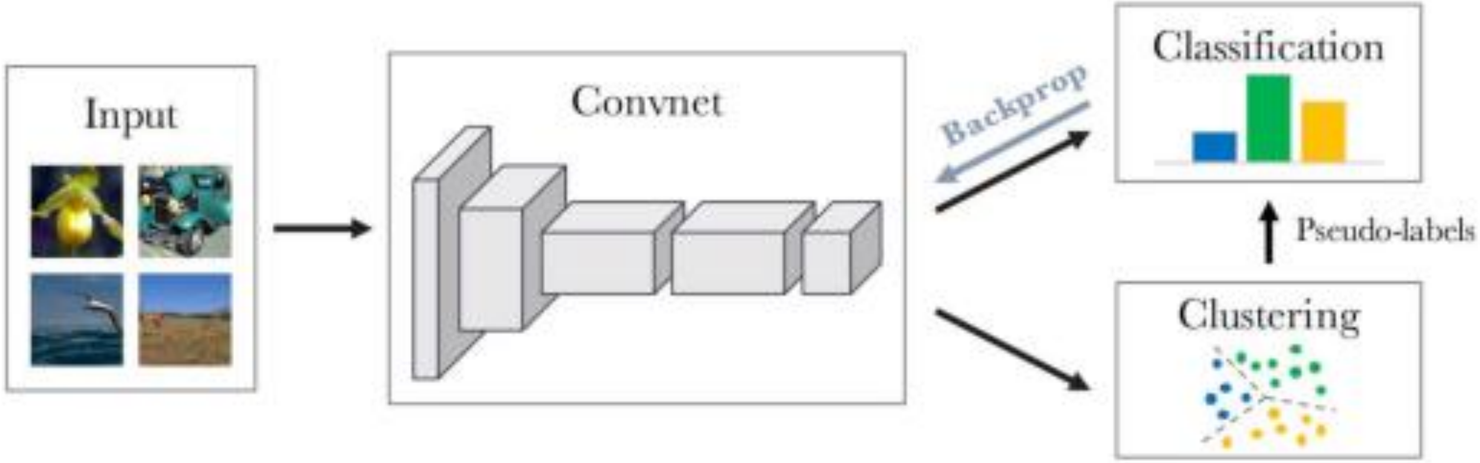
a FCN structure for colorization [1]

- Networks need to recognize objects and to group pixels of the same part together to correctly colorize each pixel

[1] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in ECCV, pp. 649–666, Springer, 2016.

Context-based image feature learning

➢ Learning with context similarity



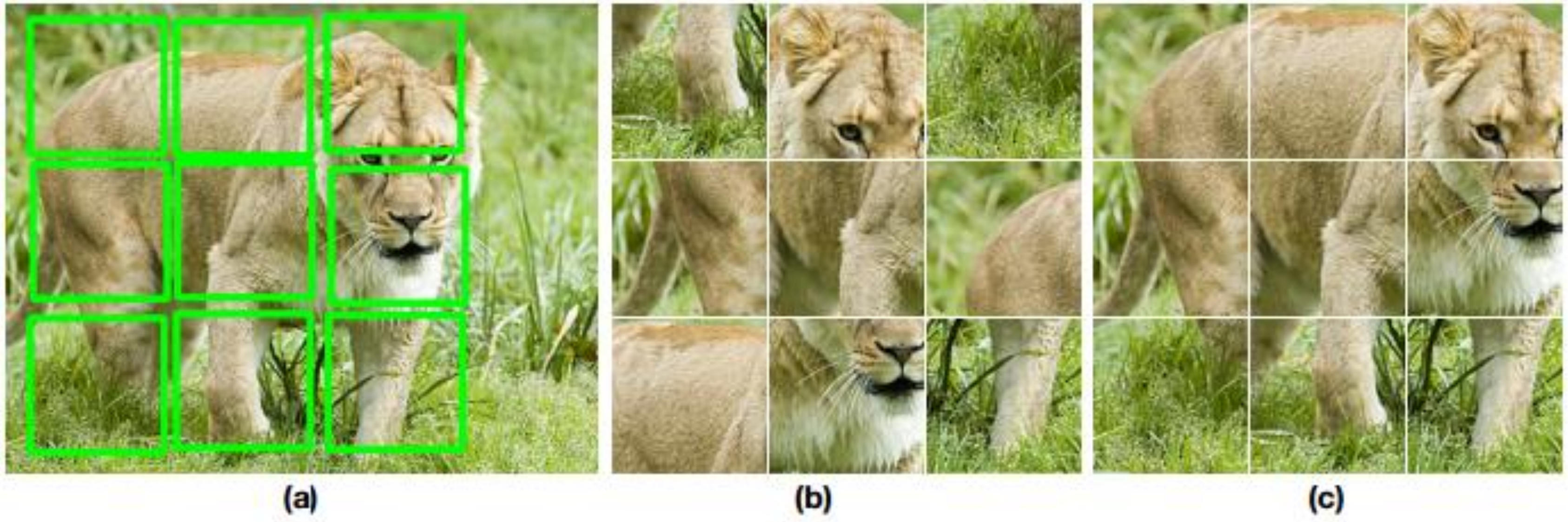
Architecture of DeepClustering [2]

- Iteratively cluster the features
- Cluster assignments are used as pseudo labels
- Initial features are generated by hand-designed features (HOG, SIFT or Fisher Vector)

[2] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in ECCV, 2018.

Context-based image feature learning

- Learning with spatial context structure
  - Predict the relative positions of two patches from same image
  - Predict the correct order of a Jigsaw puzzle
  - Recognize the rotating angles of the whole images

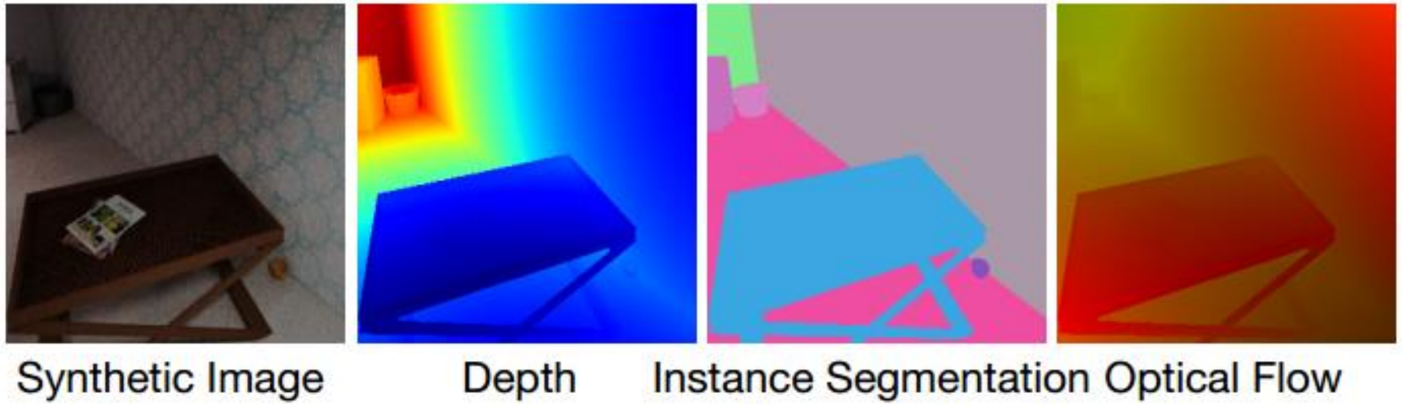


Jigsaw puzzle [3]

[3] M. Noroozi and P. Favaro, “Unsupervised learning of visual represention s by solving jigsaw puzzles,” in ECCV, 2016

Free Semantic Label-based image feature learning

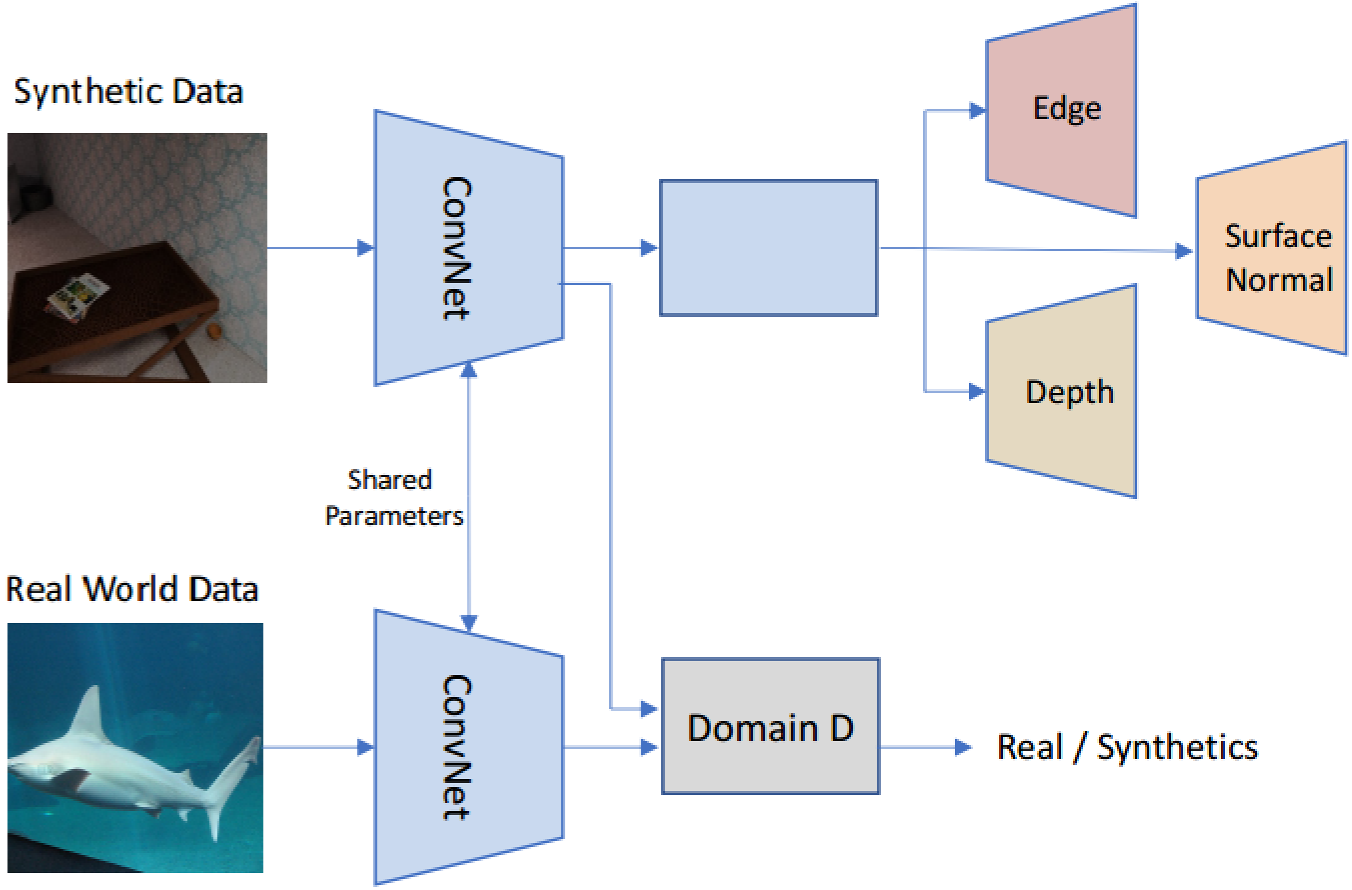
- Learning with Labels Generated by Game Engines
  - Given models of various objects and layouts of environments, game engines can render realistic images with accurate pixel-level labels



  - The domain gap between synthetic images and real-world images needs to be addressed when applied to real-world images

Free Semantic Label-based image feature learning

➢ Learning with Labels Generated by Game Engines
  • utilizing synthetic and real-world images for self-supervised feature learning



[4] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in CVPR, 2018

Free Semantic Label-based image feature learning

➢ Learning with Labels Generated by Hard-code programs

- Employing hard-code programs on images to obtain labels

- Distill knowledge from hard-code detectors, such as foreground object detection, edge detection, relative depth prediction

- **Drawback**: the semantic labels generated by hard-code detector usually are very noisy which need to specifically cope with

Video feature learning

Mehtod	SubCategory	Code	Contribution
VideoGAN [85]	Generation	✓	Forerunner of video generation with GAN
MocoGAN [86]	Generation	✓	Decomposing motion and content for video generation with GAN
TemporalGAN [144]	Generation	✓	Decomposing temporal and image generator for video generation
Video Colorization [145]	Generation	✓	Employing video colorization as the pretext task
Un-LSTM [37]	Generation	✓	Forerunner of video prediction with LSTM
ConvLSTM [146]	Generation	✓	Employing Convolutional LSTM for video prediction
MCNet [147]	Generation	✓	Disentangling motion and content for video prediction
LSTMDynamics [148]	Generation	✗	Learning by predicting long-term temporal dynamic in videos
Video Jigsaw [87]	Context	✗	Learning by jointly reasoning about spatial and temporal context
Transitive [31]	Context	✗	Learning inter and intra instance variations with a Triplet loss
3DRotNet [28]	Context	✗	Learning by recognizing rotations of video clips
CubicPuzzles [27]	Context	✗	Learning by solving video cubic puzzles
ShuffleLearn [40]	Context	✓	Employing temporal order verification as the pretext task
LSTMPermute [149]	Context	✓	Learning by temporal order verification with LSTM
OPN [39]	Context	✓	Using frame sequence order recognition as the pretext task
O3N [29]	Context	✗	Learning by identifying odd video sequences
ArrowTime [90]	Context	✓	Learning by recognizing the arrow of time in videos
TemporalCoherence [150]	Context	✗	Learning with the temporal coherence of features of frame sequence
FlowNet [151]	Cross Modal	✓	Forerunner of optical flow estimation with ConvNet
FlowNet2 [152]	Cross Modal	✓	Better architecture and better performance on optical flow estimation
UnFlow [153]	Cross Modal	✓	An unsupervised loss for optical flow estimation
CrossPixel [23]	Cross Modal	✗	Learning by predicting motion from a single image as the pretext task
CrossModel [24]	Cross Modal	✗	Optical flow and RGB correspondence verification as pretext task
AVTS [25]	Cross Modal	✗	Visual and Audio correspondence verification as pretext task
AudioVisual [26]	Cross Modal	✓	Jointly modeling visual and audio as fused multisensory representation
LookListenLearn [93]	Cross Modal	✓	Forerunner of Audio-Visual Correspondence for self-supervised learning
AmbientSound [154]	Cross Modal	✗	Predicting a statistical summary of the sound from a video frame
EgoMotion [155]	Cross Modal	✓	Learning by predicting camera motion and the scene structure from videos
LearnByMove [94]	Cross Modal	✓	Learning by predicting the camera transformation from a pairs of images
TiedEgoMotion [95]	Cross Modal	✗	Learning from ego-motor signals and video sequence
GoNet [156]	Cross Modal	✓	Jointly learning monocular depth, optical flow and ego-motion estimation from videos
DepthFlow [157]	Cross Modal	✓	Depth and optical flow learning using cross-task consistency from videos
VisualOdometry [158]	Cross Modal	✓	An unsupervised paradigm for deep visual odometry learning
ActivesStereoNet [159]	Cross Modal	✓	End-to-end self-supervised learning of depth from active stereo systems

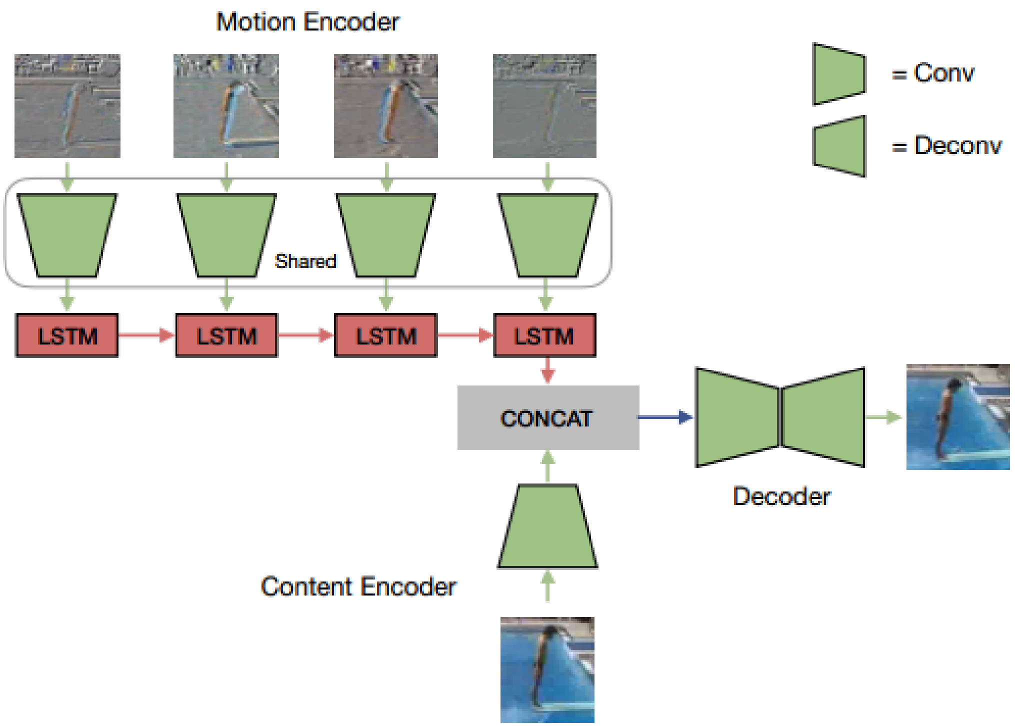
Free Semantic Label-based image feature learning

- Learning from video generation
  - Parameters of discriminator can be transferred
- Learning from video colorization
  - The color coherence between consecutive frames within a short time is a strong supervision signal
  - Given the reference RGB frame and a gray-scale image, colorize the gray-scale image
  - Another perspective is directly transform a grayscale video clip to a colorful video clip

Free Semantic Label-based image feature learning

➢ Learning from video prediction

- Predicting future frame sequences based on a limited number of frames

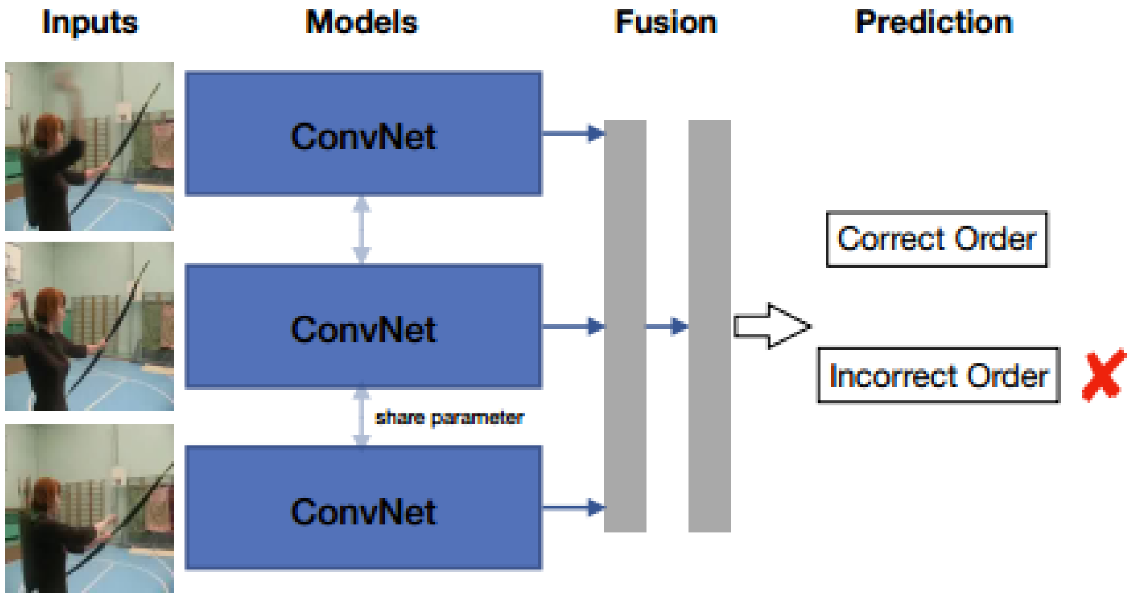


- No work has been done to study the generalization ability of features learned by video prediction

[5] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, “Decomposing motion and content for natural video sequence prediction,” in ICLR, 2017.

Temporal Context-based Learning

- ➢ Temporal order verification: correct or incorrect temporal order
- ➢ Temporal order recognition: recognize the temporal order



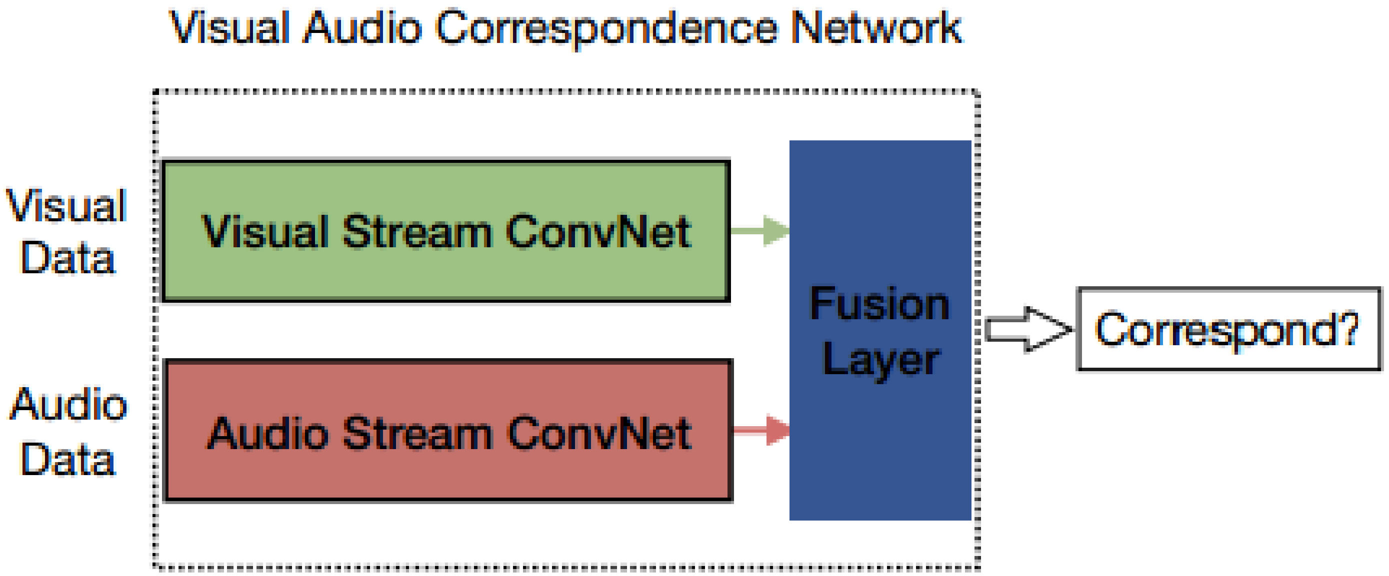
Shuffle and Learn [6]

- ➢ Frames are sampled according to the magnitude of optical flow
- ➢ Drawback: computation of optical flow is expensive and slow

[6] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: unsupervised learning using temporal order verification,” in ECCV, pp. 527–544, Springer, 2016

Cross Modal-based Learning

- Learning from RGB-Flow Correspondence
  - Optical flow estimation (e.g., FlowNets)
  - RGB and optical flow correspondence verification
- Learning from Visual-Audio Correspondence



- Ego-motion: the correspondence between visual signal and motor signal

[7] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in ICCV, pp. 609–617, IEEE, 2017

Comparison

➢ Linear classification on ImageNet and Places datasets using activations from the convolutional layers of an AlexNet as features

Method	Pretext Tasks	ImageNet conv1	ImageNet conv2	ImageNet conv3	ImageNet conv4	ImageNet conv5	Places conv1	Places conv2	Places conv3	Places conv4	Places conv5
Places labels [8]	—	—	—	—	—	—	22.1	35.1	40.2	43.3	44.6
ImageNet labels [8]	—	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random(Scratch) [8]	—	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
ColorfulColorization [18]	Generation	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
BiGAN [122]	Generation	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
SplitBrain [42]	Generation	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
ContextEncoder [19]	Context	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
ContextPrediction [41]	Context	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Jigsaw [20]	Context	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Learning2Count [130]	Context	18.0	30.6	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
DeepClustering [44]	**Context**	13.4	**32.3**	**41.0**	**39.6**	**38.2**	19.6	33.2	**39.2**	**39.8**	**34.7**

conv3 & conv4 features preform better !

- Shallow layers (conv1 & conv2) capture general low-level features
- Deep layers (conv5) capture pretext task-related features

Comparison

➢ Self-supervised image feature learning

Method	Pretext Tasks	Classification	Detection	Segmentation
ImageNet Labels [8]	—	79.9	56.8	48.0
Random(Scratch) [8]	—	57.0	44.5	30.1
ContextEncoder [19]	Generation	56.5	44.5	29.7
BiGAN [122]	Generation	60.1	46.9	35.2
ColorfulColorization [18]	Generation	65.9	46.9	35.6
SplitBrain [42]	Generation	67.1	46.7	36.0
RankVideo [38]	Context	63.1	47.2	$35.4^{\dagger}$
PredictNoise [46]	Context	65.3	49.4	$37.1^{\dagger}$
JigsawPuzzle [20]	Context	67.6	53.2	37.6
ContextPrediction [41]	Context	65.3	51.1	—
Learning2Count [130]	Context	67.7	51.4	36.6
DeepClustering [44]	**Context**	**73.7**	**55.4**	**45.1**
WatchingVideo [81]	Free Semantic Label	61.0	52.2	—
CrossDomain [30]	Free Semantic Label	68.0	52.6	—
AmbientSound [154]	Cross Modal	61.3	—	—
TiedToEgoMotion [95]	Cross Modal	—	41.7	—
EgoMotion [94]	Cross Modal	54.2	43.9	—

Comparable to supervised pre-training, especially for object detection and semantic segmentation

Comparison

➢ Self-supervised video feature learning

Method	Pretext Task	UCF101	HMDB51
Kinetics Labels* [70]	—	84.4	56.4
VideoGAN [85]	Generation	52.1	—
VideoRank [38]	Context	40.7	15.6
ShuffleLearn [40]	Context	50.9	19.8
OPN [29]	Context	56.3	22.1
RL [35]	Context	58.6	25.0
AOT [90]	Context	58.6	—
3DRotNet [28]	Context	62.9	**33.7**
CubicPuzzle* [27]	**Context**	**65.8**	**33.7**
RGB-Flow [24]	Cross Modal	59.3	27.7
PoseAction [48]	Cross Modal	55.4	23.6

Much lower than supervised pre-training, probably due to easy overfitting of 3DConvNets and the complexity of video feature learning

Future directions

- **Learning from synthetic data**: bridge the domain gap by GAN
- **Learning web data**: handle the noise in web data and their associated metadata
- **Learning spatialtemporal features from videos**: more effective pretext tasks
- **Learning with data from different sensors**: correspondence of data captured by different devices
- **Learning with multiple pretext tasks**: using different supervision signals