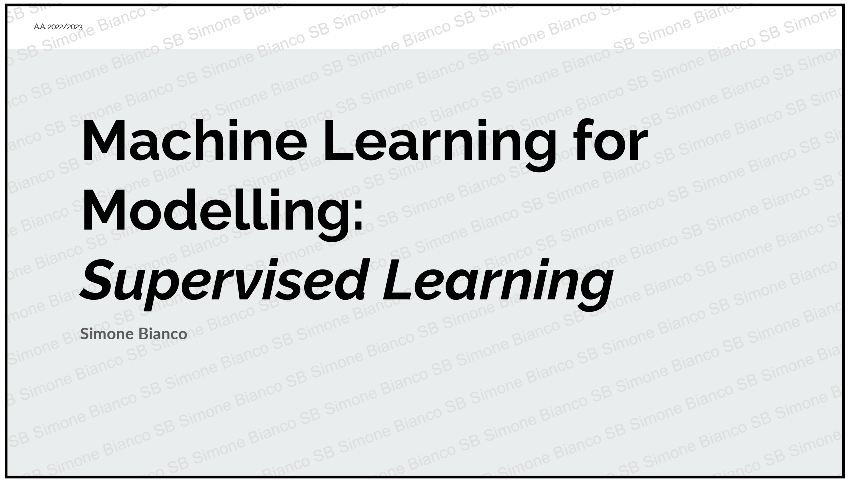
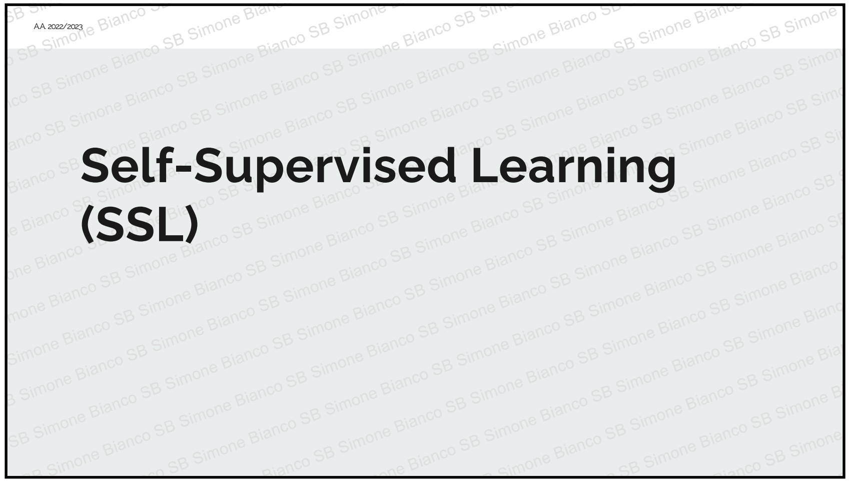


AA 2022/2023

Machine Learning for Modelling: *Supervised Learning*

Simone Bianco

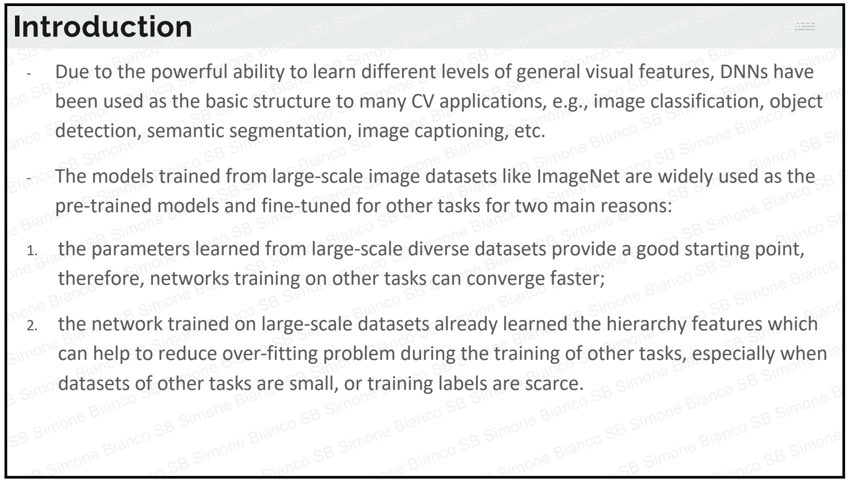
1



AA 2022/2023

Self-Supervised Learning (SSL)

2

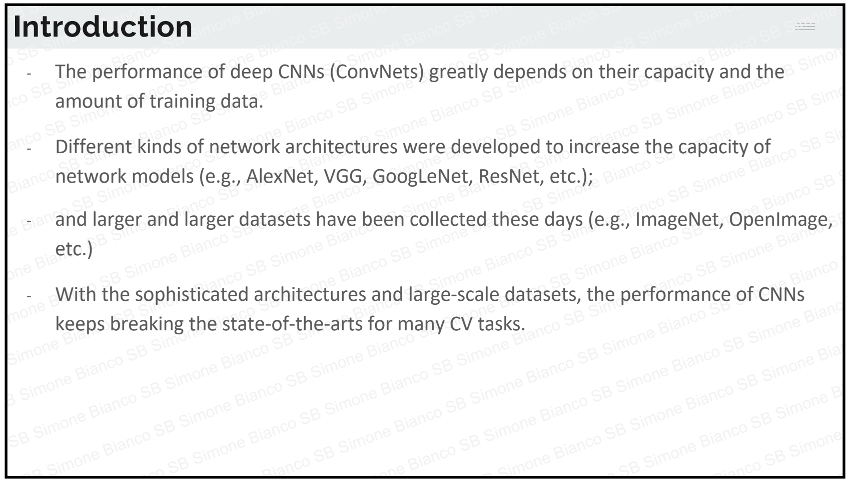


Introduction

- Due to the powerful ability to learn different levels of general visual features, DNNs have been used as the basic structure to many CV applications, e.g., image classification, object detection, semantic segmentation, image captioning, etc.
- The models trained from large-scale image datasets like ImageNet are widely used as the pre-trained models and fine-tuned for other tasks for two main reasons:

1. the parameters learned from large-scale diverse datasets provide a good starting point, therefore, networks training on other tasks can converge faster;
2. the network trained on large-scale datasets already learned the hierarchy features which can help to reduce over-fitting problem during the training of other tasks, especially when datasets of other tasks are small, or training labels are scarce.

3



Introduction

- The performance of deep CNNs (ConvNets) greatly depends on their capacity and the amount of training data.
- Different kinds of network architectures were developed to increase the capacity of network models (e.g., AlexNet, VGG, GoogLeNet, ResNet, etc.);
- and larger and larger datasets have been collected these days (e.g., ImageNet, OpenImage, etc.)
- With the sophisticated architectures and large-scale datasets, the performance of CNNs keeps breaking the state-of-the-arts for many CV tasks.

4

Introduction

- However, collection and annotation of large-scale datasets are time-consuming and expensive.
- ImageNet contains about 1.3 million labeled images covering 1,000 classes while each image is labeled by human workers with one class label.
- Compared to image datasets, collection and annotation of video datasets are more expensive due to the temporal dimension. The Kinetics dataset, which is mainly used to train CNNs for video human action recognition, consists of 500,000 videos belonging to 600 categories and each video lasts around 10 seconds.
- It took many Amazon Turk workers a lot of time to collect and annotate a dataset at such a large scale.

5

Introduction

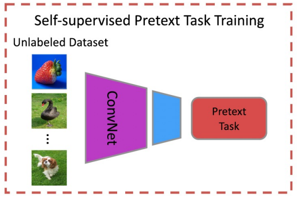
- To avoid time-consuming and expensive data annotations, many self-supervised methods were proposed to learn visual features from large-scale unlabeled images or videos without using any human annotations.
- A popular solution is to propose various pretext tasks for networks to solve. The networks can be trained by learning objective functions of the pretext tasks and in the while the features are learned through this process.
- Various pretext tasks have been proposed for self-supervised learning including colorizing grayscale images, image inpainting, playing image jigsaw puzzle, etc. The pretext tasks share two common properties:

(1) visual features of images or videos need to be captured by CNNs to solve the pretext tasks,

(2) the supervisory signal is generated from the data itself (self-supervision) by leveraging its structure.

6

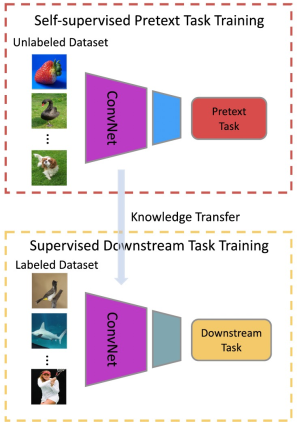
Introduction



- During the self-supervised training phase, a predefined pretext task is designed for CNNs to solve, and the pseudo labels for the pretext task are automatically generated based on some attributes of data.
- Then the CNN is trained to learn object functions of the pretext task. When trained with pretext tasks, the shallower blocks of CNN focus on the low-level general features such as corners, edges, and textures, while the deeper blocks focus on the high-level task-specific features such as objects, scenes, and object parts.
- Therefore, CNNs trained with pretext tasks can learn kernels that capture low-level features and high-level features that are helpful for other downstream tasks.

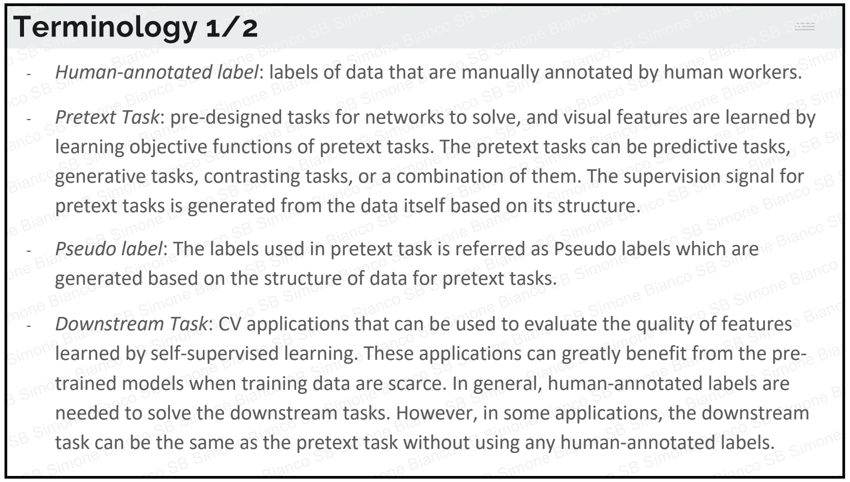
7

Introduction



- After the self-supervised training is finished, the learned visual features can be further transferred to downstream tasks (especially when only relatively small data available) as pre-trained models to improve performance and overcome over-fitting.
- Usually, visual features from only the first several layers are transferred during the supervised downstream task training phase.

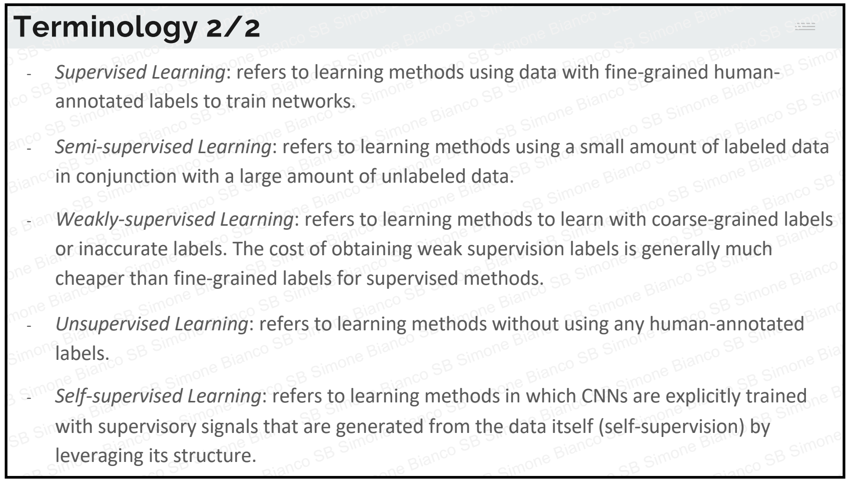
8



Terminology 1/2

- *Human-annotated label*: labels of data that are manually annotated by human workers.
- *Pretext Task*: pre-designed tasks for networks to solve, and visual features are learned by learning objective functions of pretext tasks. The pretext tasks can be predictive tasks, generative tasks, contrasting tasks, or a combination of them. The supervision signal for pretext tasks is generated from the data itself based on its structure.
- *Pseudo label*: The labels used in pretext task is referred as Pseudo labels which are generated based on the structure of data for pretext tasks.
- *Downstream Task*: CV applications that can be used to evaluate the quality of features learned by self-supervised learning. These applications can greatly benefit from the pre-trained models when training data are scarce. In general, human-annotated labels are needed to solve the downstream tasks. However, in some applications, the downstream task can be the same as the pretext task without using any human-annotated labels.

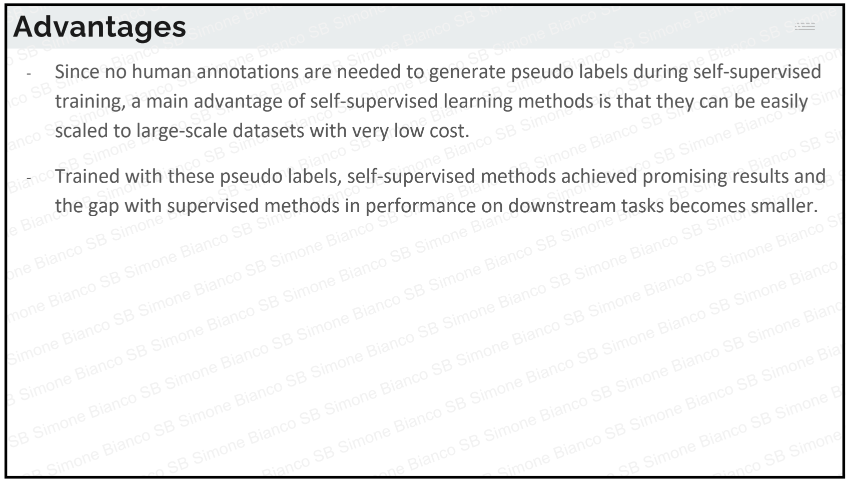
9



Terminology 2/2

- *Supervised Learning*: refers to learning methods using data with fine-grained human-annotated labels to train networks.
- *Semi-supervised Learning*: refers to learning methods using a small amount of labeled data in conjunction with a large amount of unlabeled data.
- *Weakly-supervised Learning*: refers to learning methods to learn with coarse-grained labels or inaccurate labels. The cost of obtaining weak supervision labels is generally much cheaper than fine-grained labels for supervised methods.
- *Unsupervised Learning*: refers to learning methods without using any human-annotated labels.
- *Self-supervised Learning*: refers to learning methods in which CNNs are explicitly trained with supervisory signals that are generated from the data itself (self-supervision) by leveraging its structure.

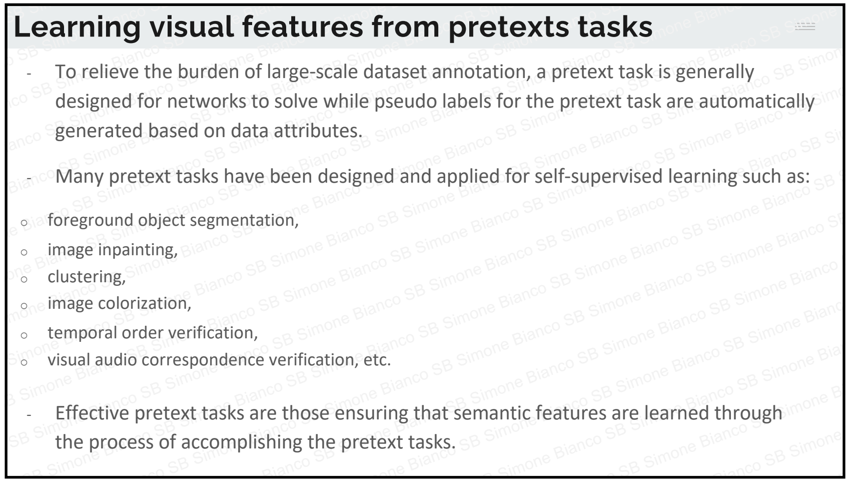
10



Advantages

- Since no human annotations are needed to generate pseudo labels during self-supervised training, a main advantage of self-supervised learning methods is that they can be easily scaled to large-scale datasets with very low cost.
- Trained with these pseudo labels, self-supervised methods achieved promising results and the gap with supervised methods in performance on downstream tasks becomes smaller.

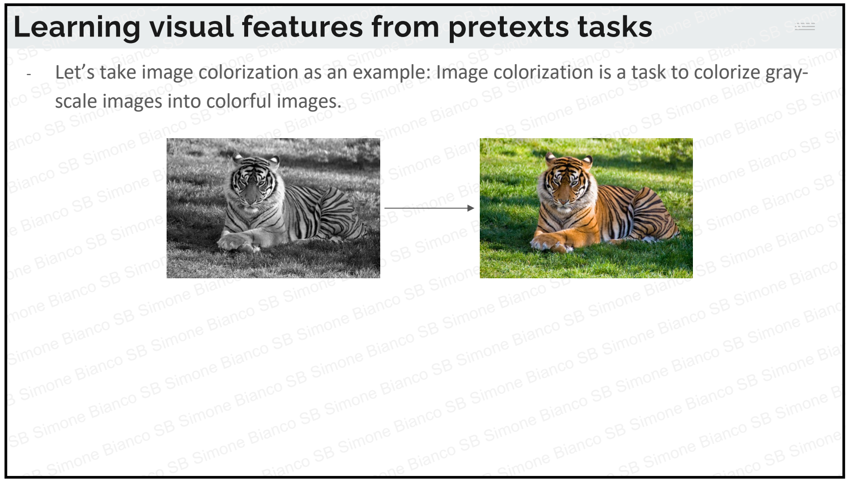
11



Learning visual features from pretexts tasks

- To relieve the burden of large-scale dataset annotation, a pretext task is generally designed for networks to solve while pseudo labels for the pretext task are automatically generated based on data attributes.
- Many pretext tasks have been designed and applied for self-supervised learning such as:
  - foreground object segmentation,
  - image inpainting,
  - clustering,
  - image colorization,
  - temporal order verification,
  - visual audio correspondence verification, etc.
- Effective pretext tasks are those ensuring that semantic features are learned through the process of accomplishing the pretext tasks.

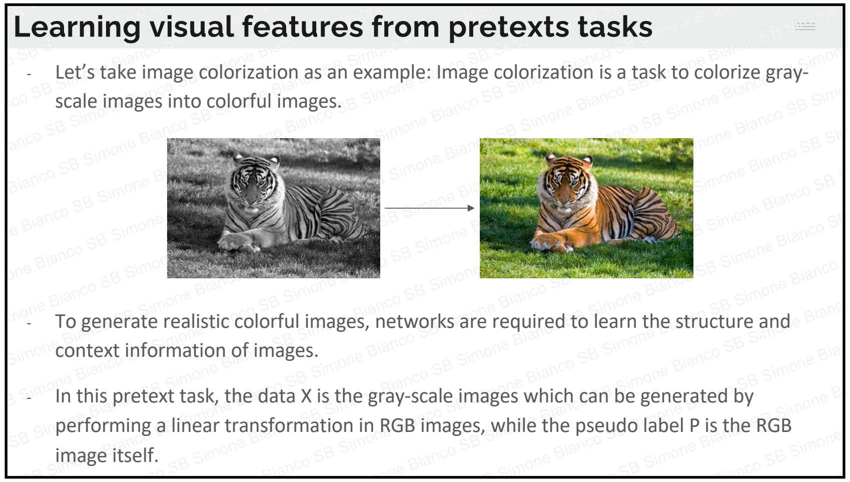
12



Learning visual features from pretexts tasks

- Let’s take image colorization as an example: Image colorization is a task to colorize gray-scale images into colorful images.

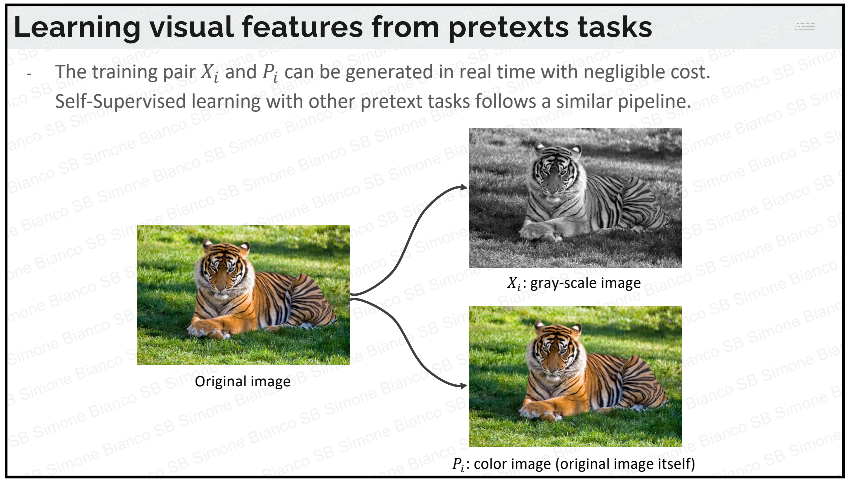
13



Learning visual features from pretexts tasks

- Let’s take image colorization as an example: Image colorization is a task to colorize gray-scale images into colorful images.
- To generate realistic colorful images, networks are required to learn the structure and context information of images.
- In this pretext task, the data X is the gray-scale images which can be generated by performing a linear transformation in RGB images, while the pseudo label P is the RGB image itself.

14



Learning visual features from pretexts tasks

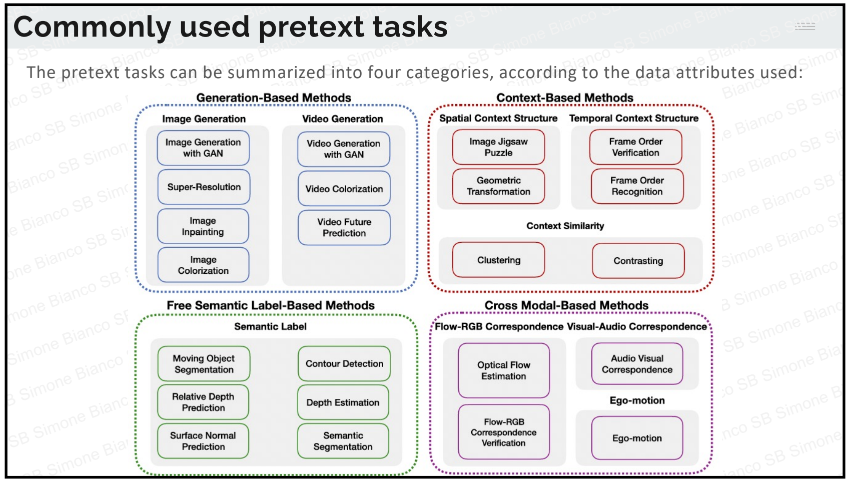
- The training pair X_i and P_i can be generated in real time with negligible cost. Self-Supervised learning with other pretext tasks follows a similar pipeline.

Original image

X_i: gray-scale image

P_i: color image (original image itself)

15



Commonly used pretext tasks

The pretext tasks can be summarized into four categories, according to the data attributes used:

Generation-Based Methods

Image Generation

Image Generation with GAN

Super-Resolution

Image Inpainting

Image Colorization

Video Generation

Video Generation with GAN

Video Colorization

Video Future Prediction

Context-Based Methods

Spatial Context Structure

Image Jigsaw Puzzle

Geometric Transformation

Temporal Context Structure

Frame Order Verification

Frame Order Recognition

Context Similarity

Clustering

Contrasting

Free Semantic Label-Based Methods

Semantic Label

Moving Object Segmentation

Relative Depth Prediction

Surface Normal Prediction

Contour Detection

Depth Estimation

Semantic Segmentation

Cross Modal-Based Methods

Flow-RGB Correspondence

Optical Flow Estimation

Flow-RGB Correspondence Verification

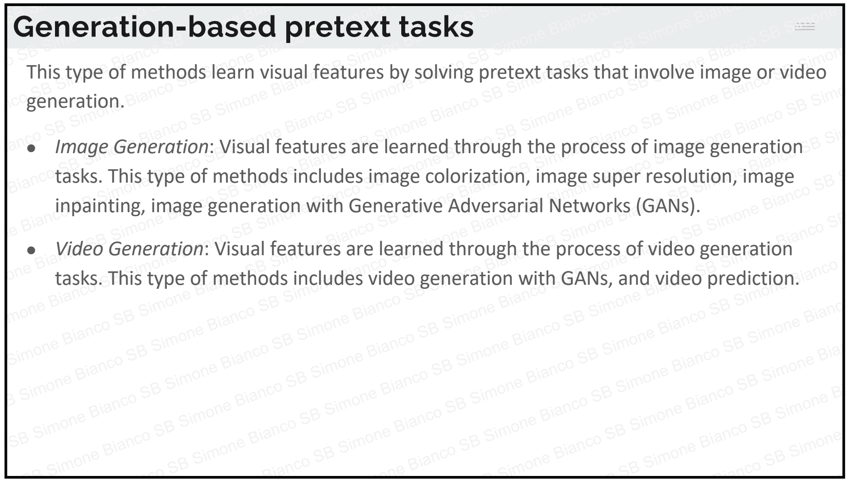
Visual-Audio Correspondence

Audio Visual Correspondence

Ego-motion

Ego-motion

16

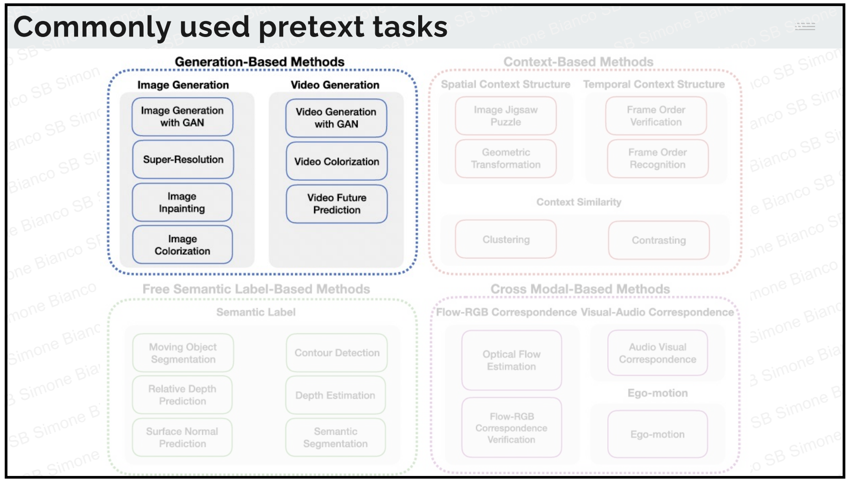


Generation-based pretext tasks

This type of methods learn visual features by solving pretext tasks that involve image or video generation.

- *Image Generation*: Visual features are learned through the process of image generation tasks. This type of methods includes image colorization, image super resolution, image inpainting, image generation with Generative Adversarial Networks (GANs).
- *Video Generation*: Visual features are learned through the process of video generation tasks. This type of methods includes video generation with GANs, and video prediction.

17



Commonly used pretext tasks

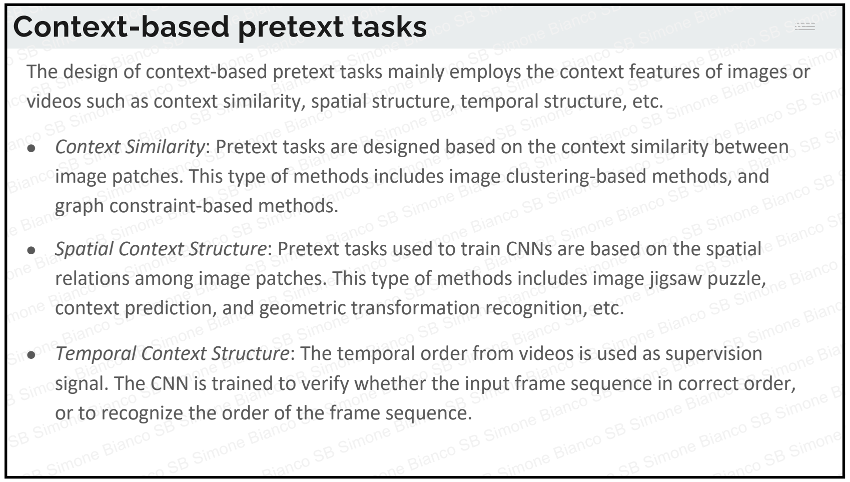
Generation-Based Methods
Image Generation
Image Generation with GAN
Super-Resolution
Image Inpainting
Image Colorization
Video Generation
Video Generation with GAN
Video Colorization
Video Future Prediction

Context-Based Methods
Spatial Context Structure
Image Jigsaw Puzzle
Geometric Transformation
Temporal Context Structure
Frame Order Verification
Frame Order Recognition
Context Similarity
Clustering
Contrasting

Free Semantic Label-Based Methods
Semantic Label
Moving Object Segmentation
Contour Detection
Relative Depth Prediction
Depth Estimation
Surface Normal Prediction
Semantic Segmentation

Cross Modal-Based Methods
Flow-RGB Correspondence
Optical Flow Estimation
Flow-RGB Correspondence Verification
Visual-Audio Correspondence
Audio Visual Correspondence
Ego-motion
Ego-motion

18

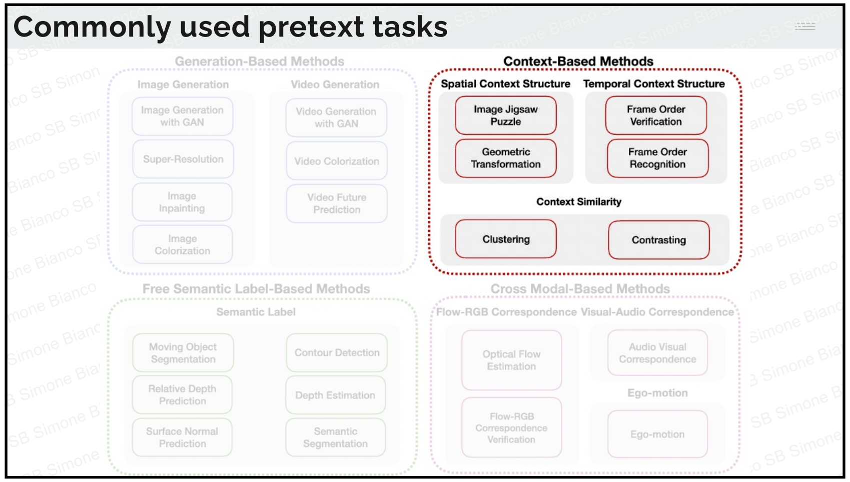


Context-based pretext tasks

The design of context-based pretext tasks mainly employs the context features of images or videos such as context similarity, spatial structure, temporal structure, etc.

- *Context Similarity*: Pretext tasks are designed based on the context similarity between image patches. This type of methods includes image clustering-based methods, and graph constraint-based methods.
- *Spatial Context Structure*: Pretext tasks used to train CNNs are based on the spatial relations among image patches. This type of methods includes image jigsaw puzzle, context prediction, and geometric transformation recognition, etc.
- *Temporal Context Structure*: The temporal order from videos is used as supervision signal. The CNN is trained to verify whether the input frame sequence in correct order, or to recognize the order of the frame sequence.

19



Commonly used pretext tasks

Generation-Based Methods
Image Generation
Image Generation with GAN
Super-Resolution
Image Inpainting
Image Colorization
Video Generation
Video Generation with GAN
Video Colorization
Video Future Prediction

Context-Based Methods
Spatial Context Structure
Image Jigsaw Puzzle
Geometric Transformation
Temporal Context Structure
Frame Order Verification
Frame Order Recognition
Context Similarity
Clustering
Contrasting

Free Semantic Label-Based Methods
Semantic Label
Moving Object Segmentation
Contour Detection
Relative Depth Prediction
Depth Estimation
Surface Normal Prediction
Semantic Segmentation

Cross Modal-Based Methods
Flow-RGB Correspondence
Optical Flow Estimation
Flow-RGB Correspondence Verification
Visual-Audio Correspondence
Audio Visual Correspondence
Ego-motion
Ego-motion

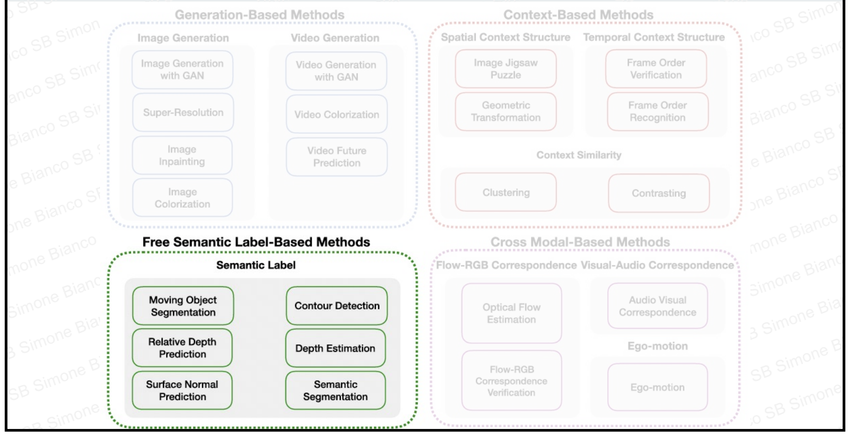
20

Free semantic label-based pretext tasks

- This type of pretext tasks train networks with automatically generated semantic labels.
- The labels are generated by traditional hard-code algorithms or by game engines.
- Since no human-annotations are involved through the design of hard-code algorithms, the detectors can be used to generate labels for self-supervised training.
- Strictly speaking, the methods based on data generated by game engines do/should not belong to the self-supervised learning methods since human intervention is needed during the data generation process. However, some recent work treat them as self-supervised learning methods.
- We also include this type of methods here such as moving object segmentation, contour detection, relative depth prediction, etc.

21

Commonly used pretext tasks



Generation-Based Methods

Image Generation: Image Generation with GAN; Super-Resolution; Image Inpainting; Image Colorization

Video Generation: Video Generation with GAN; Video Colorization; Video Future Prediction

Context-Based Methods

Spatial Context Structure: Image Jigsaw Puzzle; Geometric Transformation

Temporal Context Structure: Frame Order Verification; Frame Order Recognition

Context Similarity: Clustering; Contrasting

Free Semantic Label-Based Methods

Semantic Label: Moving Object Segmentation; Contour Detection; Relative Depth Prediction; Depth Estimation; Surface Normal Prediction; Semantic Segmentation

Cross Modal-Based Methods

Flow-RGB Correspondence: Optical Flow Estimation; Flow-RGB Correspondence Verification

Visual-Audio Correspondence: Audio Visual Correspondence

Ego-motion: Ego-motion

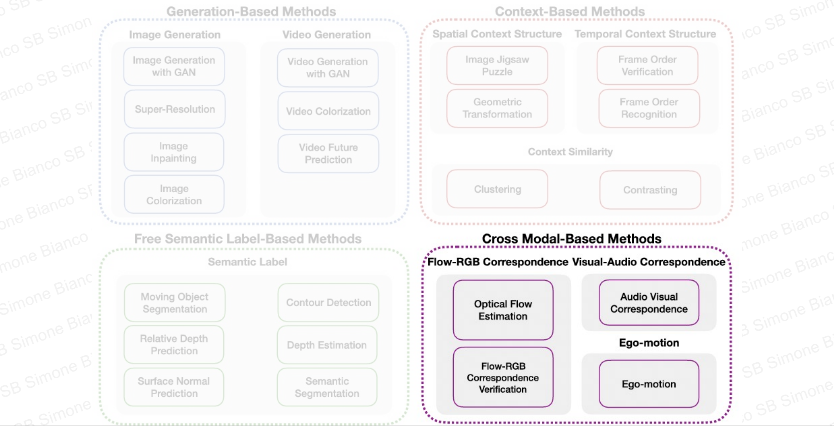
22

Cross modal-based pretext tasks

- This type of pretext tasks trains CNNs to verify whether two different channels of input data are corresponding to each other such as Visual-Audio Correspondence Verification, RGB-Flow Correspondence Verification, Contrasting (i.e., maximizing agreement between differently augmented views of the same data example), and egomotion.

23

Commonly used pretext tasks



Generation-Based Methods

Image Generation: Image Generation with GAN; Super-Resolution; Image Inpainting; Image Colorization

Video Generation: Video Generation with GAN; Video Colorization; Video Future Prediction

Context-Based Methods

Spatial Context Structure: Image Jigsaw Puzzle; Geometric Transformation

Temporal Context Structure: Frame Order Verification; Frame Order Recognition

Context Similarity: Clustering; Contrasting

Free Semantic Label-Based Methods

Semantic Label: Moving Object Segmentation; Contour Detection; Relative Depth Prediction; Depth Estimation; Surface Normal Prediction; Semantic Segmentation

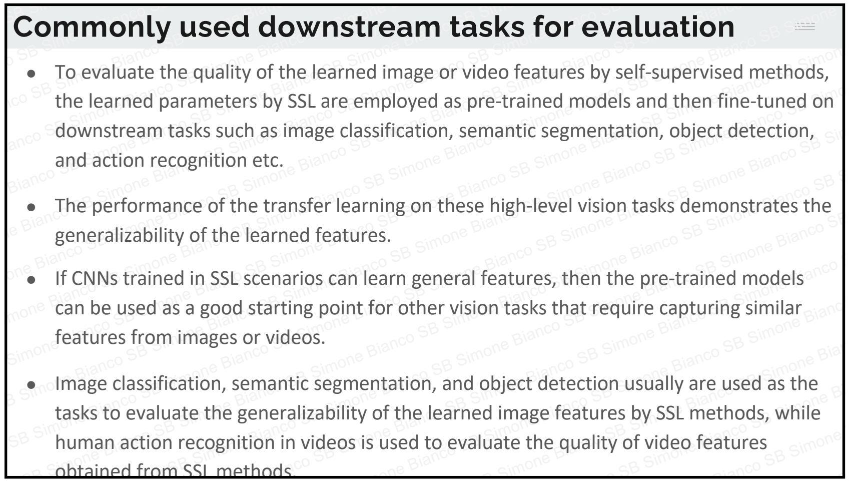
Cross Modal-Based Methods

Flow-RGB Correspondence: Optical Flow Estimation; Flow-RGB Correspondence Verification

Visual-Audio Correspondence: Audio Visual Correspondence

Ego-motion: Ego-motion

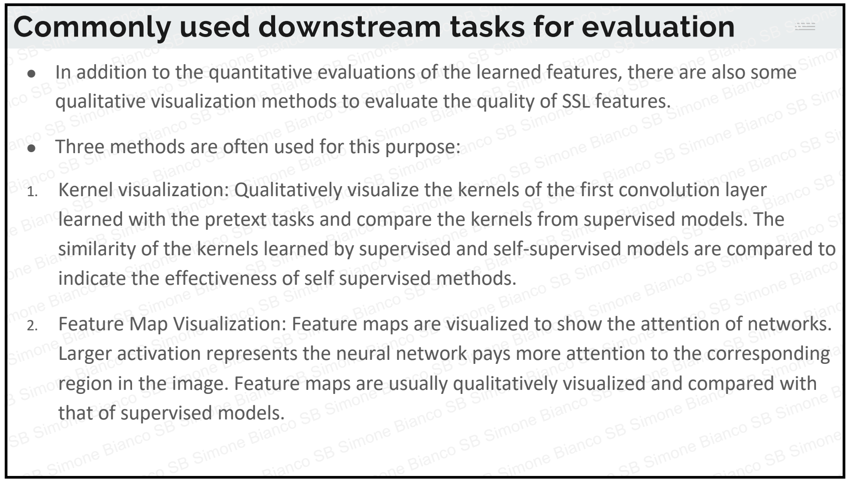
24



Commonly used downstream tasks for evaluation

- To evaluate the quality of the learned image or video features by self-supervised methods, the learned parameters by SSL are employed as pre-trained models and then fine-tuned on downstream tasks such as image classification, semantic segmentation, object detection, and action recognition etc.
- The performance of the transfer learning on these high-level vision tasks demonstrates the generalizability of the learned features.
- If CNNs trained in SSL scenarios can learn general features, then the pre-trained models can be used as a good starting point for other vision tasks that require capturing similar features from images or videos.
- Image classification, semantic segmentation, and object detection usually are used as the tasks to evaluate the generalizability of the learned image features by SSL methods, while human action recognition in videos is used to evaluate the quality of video features obtained from SSL methods

25

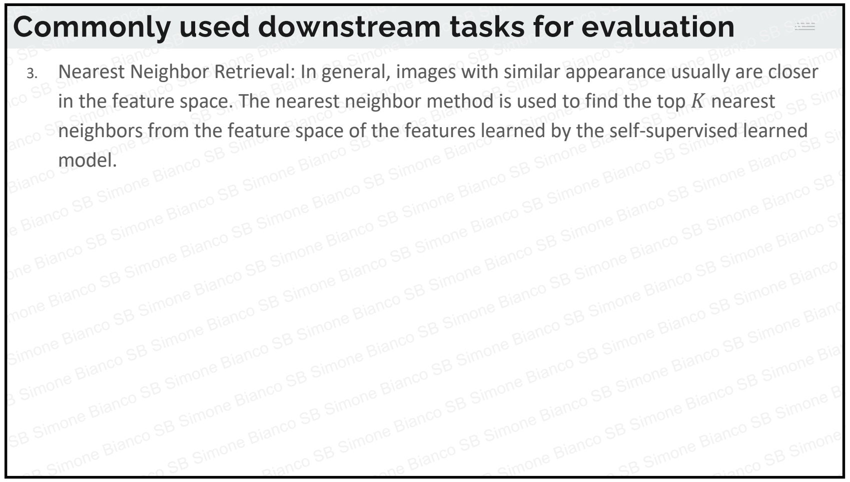


Commonly used downstream tasks for evaluation

- In addition to the quantitative evaluations of the learned features, there are also some qualitative visualization methods to evaluate the quality of SSL features.
- Three methods are often used for this purpose:

1. Kernel visualization: Qualitatively visualize the kernels of the first convolution layer learned with the pretext tasks and compare the kernels from supervised models. The similarity of the kernels learned by supervised and self-supervised models are compared to indicate the effectiveness of self supervised methods.
2. Feature Map Visualization: Feature maps are visualized to show the attention of networks. Larger activation represents the neural network pays more attention to the corresponding region in the image. Feature maps are usually qualitatively visualized and compared with that of supervised models.

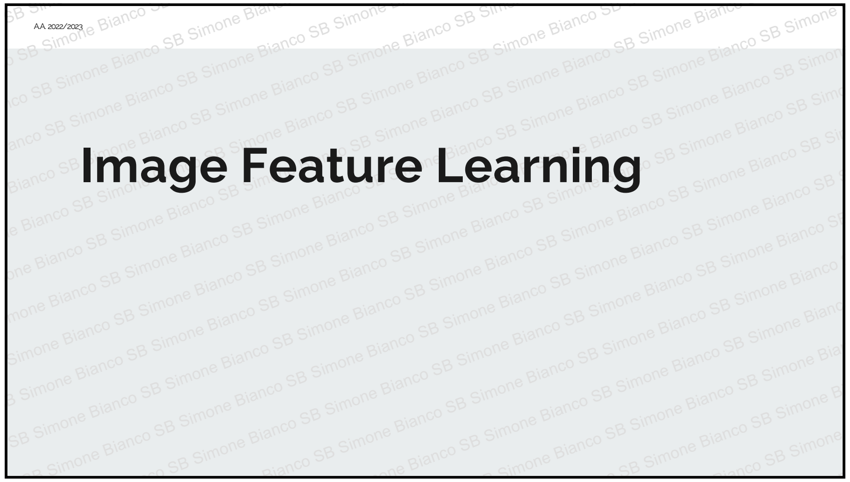
26



Commonly used downstream tasks for evaluation

3. Nearest Neighbor Retrieval: In general, images with similar appearance usually are closer in the feature space. The nearest neighbor method is used to find the top K nearest neighbors from the feature space of the features learned by the self-supervised learned model.

27



A.A. 2022/2023

Image Feature Learning

28

AA 2022/2023

Generation-based Image Feature Learning

29

Generation-based image feature learning

- Generation-based self-supervised methods for learning image features involve the process of generating images including image generation with GAN (to generate fake images), super-resolution (to generate high-resolution images), image inpainting (to predict missing image regions), and image colorization (to colorize gray-scale images into colorful images).
- For these tasks, pseudo training labels P usually are the images themselves and no human-annotated labels are needed during training, therefore, these methods belong to self-supervised learning methods.
- The pioneer work about the image generation-based methods is the Autoencoder which learns to compress an image into a low-dimension vector and then is uncompressed into an image that is close to the original image with a bunch of layers.
- With an auto-encoder, networks can reduce the dimension of an image into a lower-dimensional vector that contains the main information of the original image.

30

Generation-based image feature learning

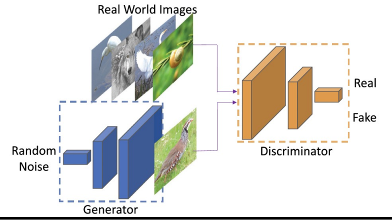
- Variational Autoencoder (VAE) is an improved version of Autoencoder which estimates the Probability Density Function (PDF) of the training data. The current image generation-based methods follow a similar idea but with different pipelines to learn visual features through the process of image generation.

31

Image generation with GAN

- Generative Adversarial Network (GAN) is a type of deep generative model that was proposed by Goodfellow et al.
- A GAN model generally consists of two kinds of networks:

1) a generator, which is to generate images from latent vectors,
2) and a discriminator, which is to distinguish whether the input image is generated by the generator.



32

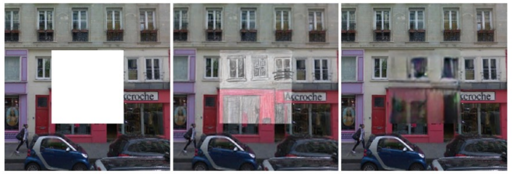
Image generation with GAN

- By playing the two-player game (that is where the *Adversarial* terms comes from), the discriminator forces the generator to generate realistic images, while the generator forces the discriminator to improve its differentiation ability.
- During the training, the two networks are competing against with each other and make each other stronger.
- The generator is trained to map any latent vector sampled from latent space into an image, while the discriminator is forced to distinguish whether the image from the real data distribution or generated data distribution.
- Therefore, the discriminator is required to capture the semantic features from images to accomplish the task.
- The parameters of the discriminator can server as the pre-trained model for other computer vision tasks.

33

Image generation with Inpainting

- Image inpainting is a task of predicting arbitrary missing regions based on the rest of an image



(a) Input Context (b) Human Artist (c) Network Prediction

- To correctly predict missing regions, networks are required to learn the common knowledge including the color and structure of the common objects.
- Only by knowing this knowledge, networks are able to infer missing regions based on the rest part of the image.

34

Image generation with Inpainting

- Usually, there are two kinds of networks:

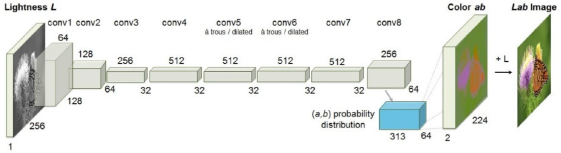
1) a generator network is to generate the missing region with the pixel-wise reconstruction loss
2) and a discriminator network is to distinguish whether the input image is real with an adversarial loss.

- With the adversarial loss, the network is able to generate sharper and realistic hypothesis for the missing image region.
- Both the two kinds of networks are able to learn the semantic features from images and can be transferred to other computer vision tasks.

35

Image generation with Colorization

- Image colorization is a task of predicting a plausible color version of the photograph given a gray-scale photograph as input.



- To correctly colorize each pixel, networks need to recognize objects and to group pixels of the same part together.
- Therefore, visual features can be learned in the process of accomplishing this task.

36

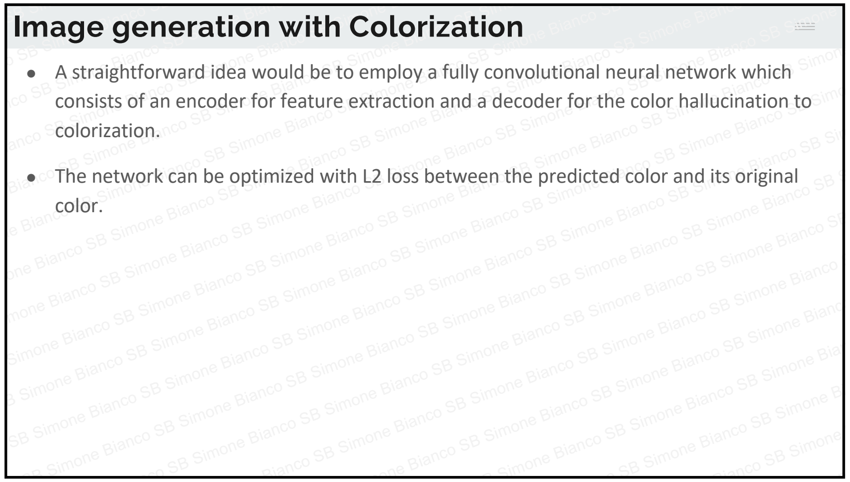
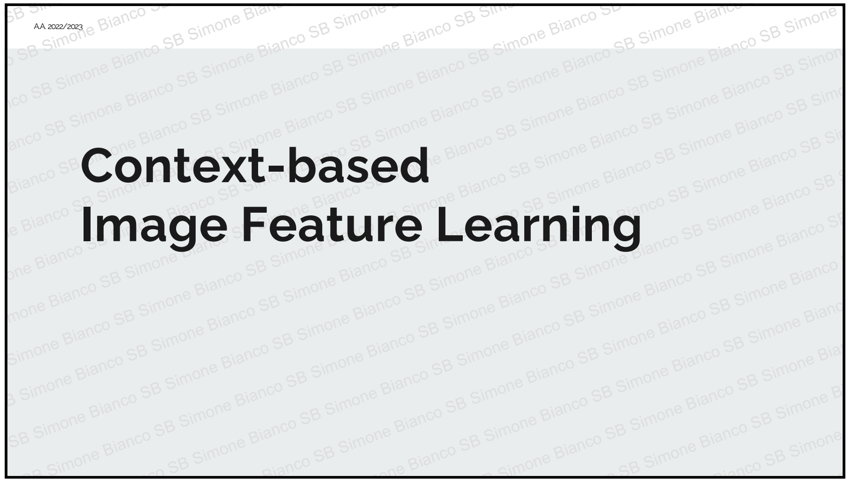


Image generation with Colorization

- A straightforward idea would be to employ a fully convolutional neural network which consists of an encoder for feature extraction and a decoder for the color hallucination to colorization.
- The network can be optimized with L2 loss between the predicted color and its original color.

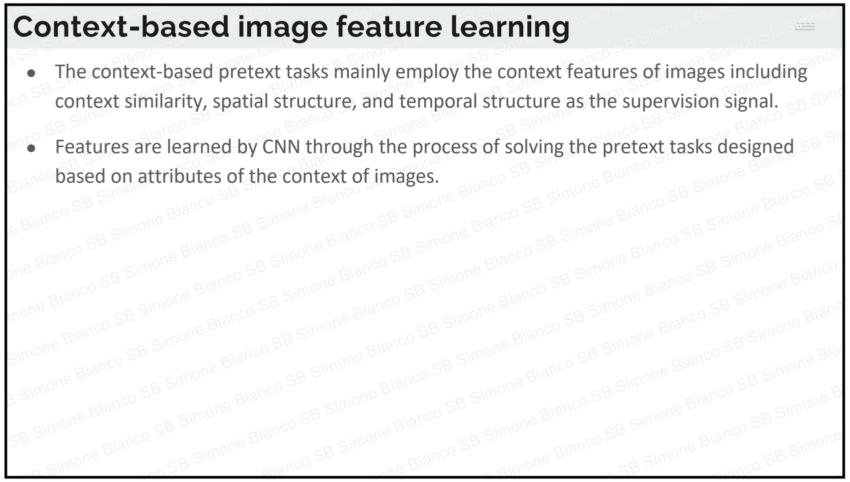
37



AA 2022/2023

Context-based Image Feature Learning

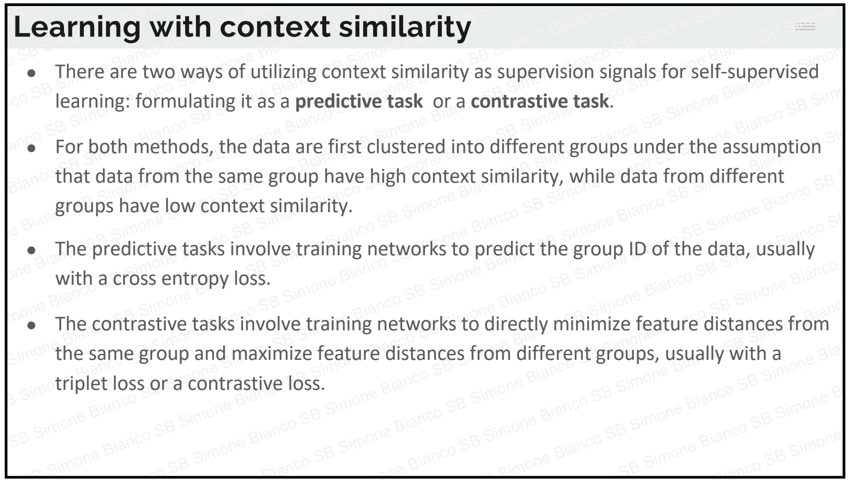
38



Context-based image feature learning

- The context-based pretext tasks mainly employ the context features of images including context similarity, spatial structure, and temporal structure as the supervision signal.
- Features are learned by CNN through the process of solving the pretext tasks designed based on attributes of the context of images.

39



Learning with context similarity

- There are two ways of utilizing context similarity as supervision signals for self-supervised learning: formulating it as a **predictive task** or a **contrastive task**.
- For both methods, the data are first clustered into different groups under the assumption that data from the same group have high context similarity, while data from different groups have low context similarity.
- The predictive tasks involve training networks to predict the group ID of the data, usually with a cross entropy loss.
- The contrastive tasks involve training networks to directly minimize feature distances from the same group and maximize feature distances from different groups, usually with a triplet loss or a contrastive loss.

40

Learning with context similarity: predictive

- Clustering is a method of grouping sets of similar data in the same clusters. In SSL, the clustering methods are mainly employed as a tool to cluster image data.
- A naive method would be to cluster the image data based on the hand-designed feature such as SIFT followed by BoW.
- After the clustering, several clusters are obtained while the image within one cluster has a smaller distance in feature space and images from different clusters have a larger distance in feature space. The smaller the distance in feature space, the more similar the image in the appearance in the RGB space.
- Then a CNN can be trained to classify the data by using the cluster assignment as the pseudo class label.
- To accomplish this task, the CNN needs to learn the invariance within one class and the variance among different classes. Therefore, the CNN learns semantic meaning of images.

41

Learning with context similarity: contrastive

- Another way of leveraging context similarly for self-supervised image feature learning is by contrasting.
- The general idea of the contrastive SSL is to train networks to maximum agreement of different views of same scene while minimizing agreement of views from different scenes.
- The recent state-of-the-art method is SimCLR which learns features by contrasting images after a composition of data augmentations.
- The positive pairs are constructed by sampling two images after applying different augmentation techniques for the same image, while negative pairs include two different images.
- This method significantly outperforms other SSL methods on ImageNet dataset.

42

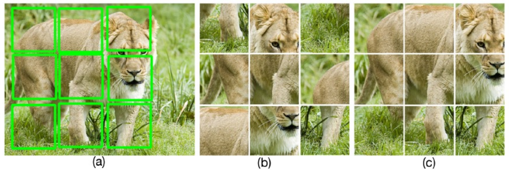
Learning with spatial context structure

- Images contain rich spatial context information such as the relative positions among different patches from an image, which can be used to design the pretext task for SSL.
- The pretext task can be to predict the relative positions of two patches from same image, or to recognize the order of a shuffled sequence of patches from same image.
- The context of full images can also be used as a supervision signal to design pretext tasks such as to recognize the rotating angles of the whole images.
- To accomplish these pretext tasks, CNNs need to learn spatial context information such as the shape of the objects and the relative positions of different parts of an object.

43

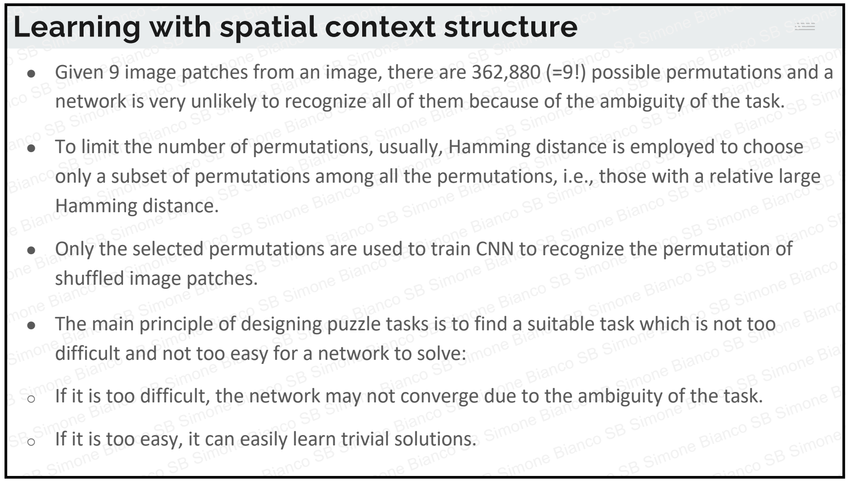
Learning with spatial context structure

- Following this idea, more methods are proposed to learn image features by solving more difficult spatial puzzles.
- One typical work attempted to solve an image Jigsaw puzzle with CNNs.



- The shuffled image patches are fed to the network which is trained to recognize the correct spatial locations of the input patches by learning spatial context structures of images such as object color, structure, and high-level semantic information.

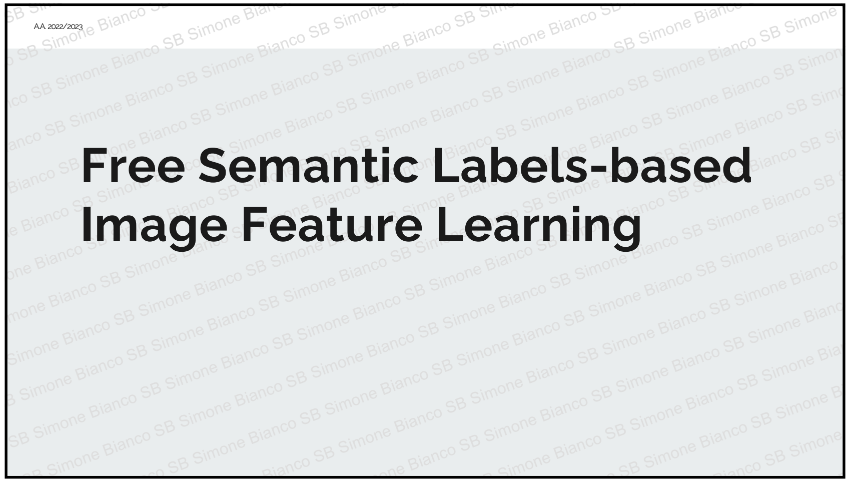
44



Learning with spatial context structure

- Given 9 image patches from an image, there are 362,880 (=9!) possible permutations and a network is very unlikely to recognize all of them because of the ambiguity of the task.
- To limit the number of permutations, usually, Hamming distance is employed to choose only a subset of permutations among all the permutations, i.e., those with a relative large Hamming distance.
- Only the selected permutations are used to train CNN to recognize the permutation of shuffled image patches.
- The main principle of designing puzzle tasks is to find a suitable task which is not too difficult and not too easy for a network to solve:
  - If it is too difficult, the network may not converge due to the ambiguity of the task.
  - If it is too easy, it can easily learn trivial solutions.

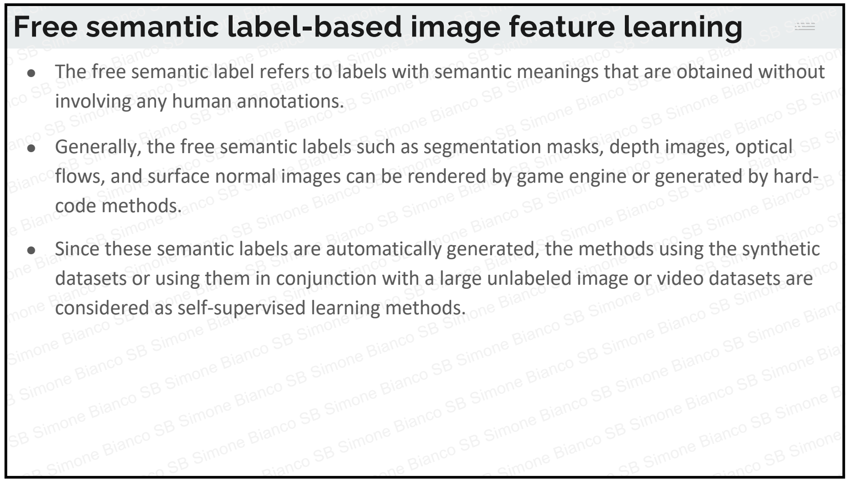
45



AA 2022/2023

Free Semantic Labels-based Image Feature Learning

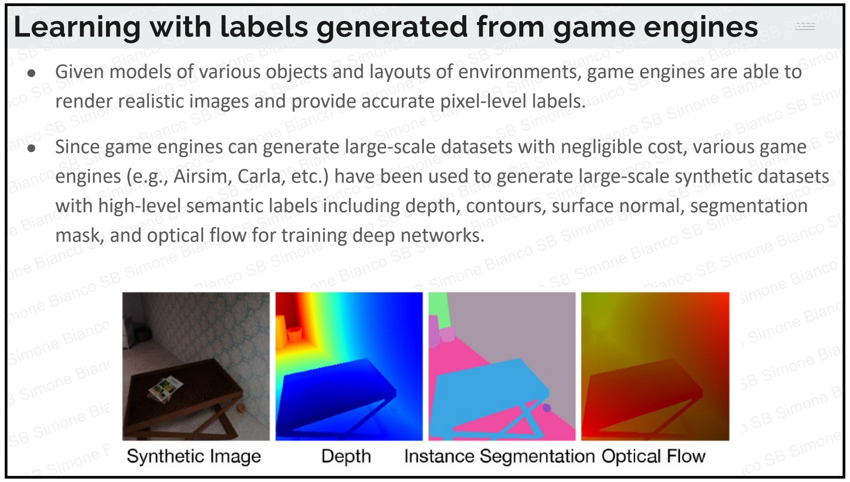
46



Free semantic label-based image feature learning

- The free semantic label refers to labels with semantic meanings that are obtained without involving any human annotations.
- Generally, the free semantic labels such as segmentation masks, depth images, optical flows, and surface normal images can be rendered by game engine or generated by hard-code methods.
- Since these semantic labels are automatically generated, the methods using the synthetic datasets or using them in conjunction with a large unlabeled image or video datasets are considered as self-supervised learning methods.

47



Learning with labels generated from game engines

- Given models of various objects and layouts of environments, game engines are able to render realistic images and provide accurate pixel-level labels.
- Since game engines can generate large-scale datasets with negligible cost, various game engines (e.g., Airsim, Carla, etc.) have been used to generate large-scale synthetic datasets with high-level semantic labels including depth, contours, surface normal, segmentation mask, and optical flow for training deep networks.

48

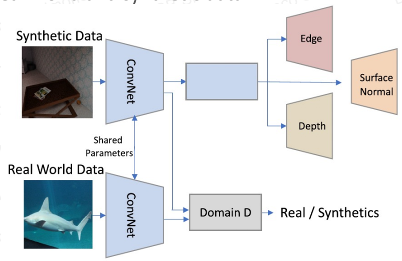
Learning with labels generated from game engines

- Game engines can generate realistic images with accurate pixel-level labels with very low cost.
- However, due to the domain gap between synthetic and real-world images, the CNNs purely trained on synthetic images cannot be directly applied to real-world images.
- To utilize synthetic datasets for self-supervised feature learning, the domain gap needs to be explicitly bridged.
- In this way, the CNN trained with the semantic labels of the synthetic dataset can be effectively applied to real-world images.

49

Learning with labels generated from game engines

- To overcome the problem, unsupervised feature space domain adaptation methods based on adversarial learning have been proposed.
- The CNN predicts surface normal, depth, and instance contour for the synthetic images and a discriminator network D is employed to minimize the difference of feature space domains between real-world and synthetic data.



50

Learning with labels generated by hard-code programs

- Applying hard-code programs is another way to automatically generate semantic labels such as salience, foreground masks, contours, depth for images and videos.
- With these methods, very large-scale datasets with generated semantic labels can be used for self-supervised feature learning.
- This type of methods generally has two steps:

1) label generation by employing hard-code programs on images or videos to obtain labels,
2) train CNNs with the generated labels.

- Various hard-code programs have been applied to generate labels for self-supervised learning methods including methods for foreground object segmentation, edge detection, and relative depth prediction.

51

Learning with labels generated by hard-code programs

- No matter what kind of labels used to train CNNs, the general idea of this type of methods is to distill knowledge from the hard-code detector.
- The hard-code detector can be edge detector, salience detector, relative depth detector, etc.
- Compared to other self-supervised learning methods, the supervision signal in these pretext tasks are semantic labels, which can directly drive the CNN to learn semantic features.
- However, one drawback is that the semantic labels generated by hard-code detector usually are very noisy which need to be specifically coped with.

52

AA 2022/2023

Cross Modal-based Learning

53

Cross modal-based learning

- Cross modal-based learning methods usually learn features from the correspondence of multiple data streams including RGB frame sequence, optical flow sequence, audio data, and camera pose.
- In addition to rich temporal and spatial information in videos, optical flow sequence can be generated to specifically indicate the motion in videos, and the difference of frames can be computed with negligible time and space-time complexity to indicate the boundary of the moving objects.
- Similarly, audio data also provide a useful hint about the content of videos.

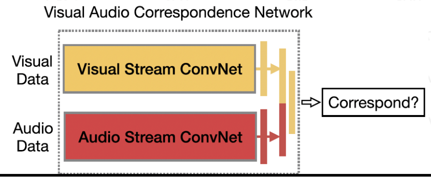
54

Cross modal-based learning

- Based on the type of data used, these methods fall into three groups:

  - methods that learn features by using the RGB and optical flow correspondence
  - methods that learn features by utilizing the video and audio correspondence
  - ego-motion that learn by utilizing the correspondence between egocentric video and egomotor sensor signals.

- Usually, the network is trained to recognize if the two kinds of input data are corresponding to each other, or is trained to learn the transformation between different modalities

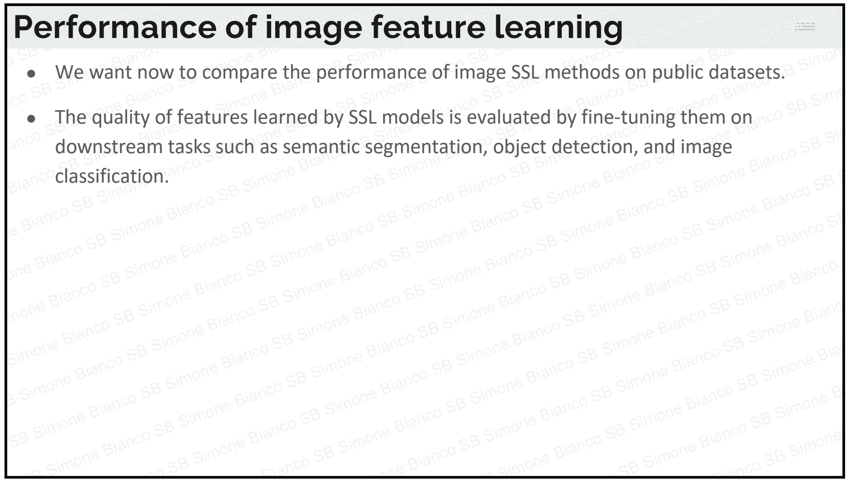


55

AA 2022/2023

Performance comparison

56



Performance of image feature learning

- We want now to compare the performance of image SSL methods on public datasets.
- The quality of features learned by SSL models is evaluated by fine-tuning them on downstream tasks such as semantic segmentation, object detection, and image classification.

57

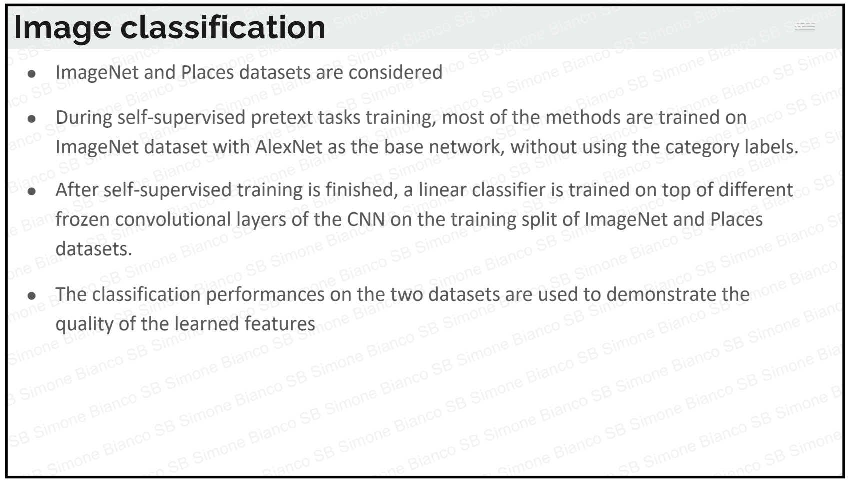


Image classification

- ImageNet and Places datasets are considered
- During self-supervised pretext tasks training, most of the methods are trained on ImageNet dataset with AlexNet as the base network, without using the category labels.
- After self-supervised training is finished, a linear classifier is trained on top of different frozen convolutional layers of the CNN on the training split of ImageNet and Places datasets.
- The classification performances on the two datasets are used to demonstrate the quality of the learned features

58

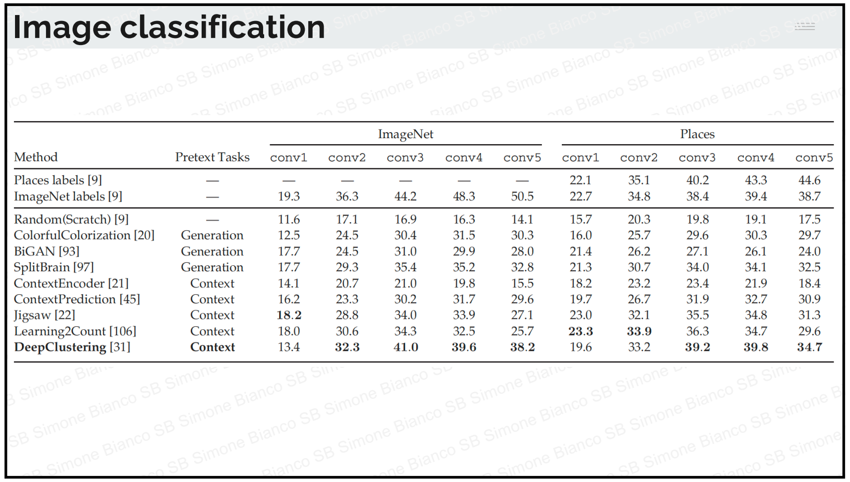


Image classification

		ImageNet					Places				
Method	Pretext Tasks	conv1	conv2	conv3	conv4	conv5	conv1	conv2	conv3	conv4	conv5
Places labels [9]	—	—	—	—	—	—	22.1	35.1	40.2	43.3	44.6
ImageNet labels [9]	—	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random(Scratch) [9]	—	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
ColorfulColorization [20]	Generation	12.5	24.5	30.4	31.5	30.3	16.0	25.7	29.6	30.3	29.7
BiGAN [93]	Generation	17.7	24.5	31.0	29.9	28.0	21.4	26.2	27.1	26.1	24.0
SplitBrain [97]	Generation	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34.0	34.1	32.5
ContextEncoder [21]	Context	14.1	20.7	21.0	19.8	15.5	18.2	23.2	23.4	21.9	18.4
ContextPrediction [45]	Context	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Jigsaw [22]	Context	**18.2**	28.8	34.0	33.9	27.1	23.0	32.1	35.5	34.8	31.3
Learning2Count [106]	Context	18.0	30.6	34.3	32.5	25.7	**23.3**	**33.9**	36.3	34.7	29.6
DeepClustering [31]	**Context**	13.4	**32.3**	**41.0**	**39.6**	**38.2**	19.6	33.2	**39.2**	**39.8**	**34.7**

59

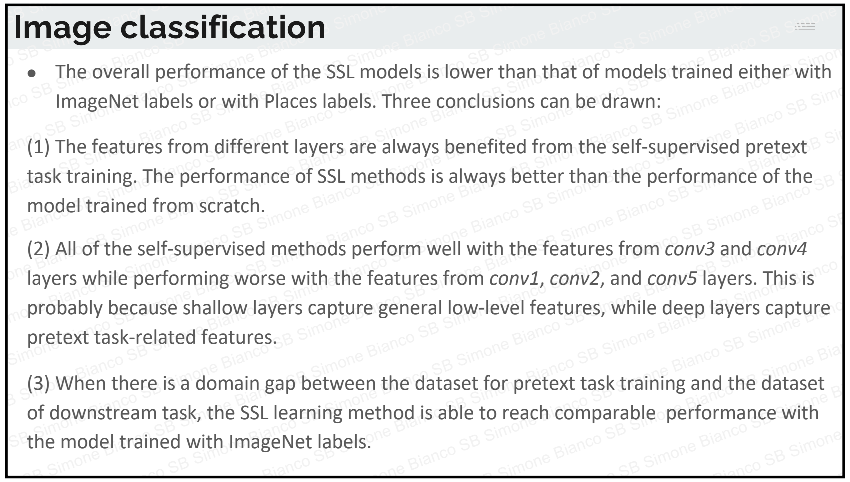


Image classification

- The overall performance of the SSL models is lower than that of models trained either with ImageNet labels or with Places labels. Three conclusions can be drawn:

(1) The features from different layers are always benefited from the self-supervised pretext task training. The performance of SSL methods is always better than the performance of the model trained from scratch.

(2) All of the self-supervised methods perform well with the features from *conv3* and *conv4* layers while performing worse with the features from *conv1*, *conv2*, and *conv5* layers. This is probably because shallow layers capture general low-level features, while deep layers capture pretext task-related features.

(3) When there is a domain gap between the dataset for pretext task training and the dataset of downstream task, the SSL learning method is able to reach comparable performance with the model trained with ImageNet labels.

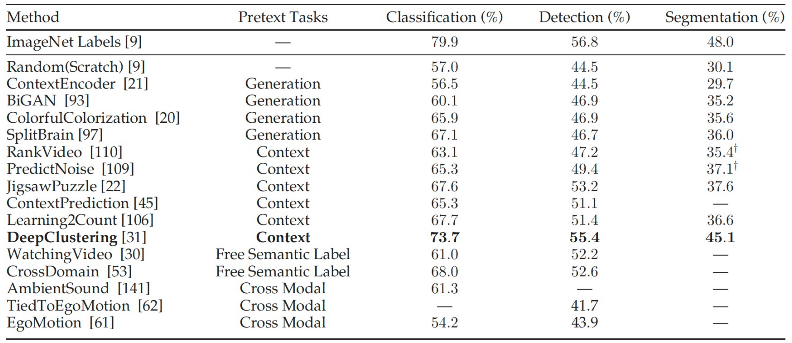
60

Image classification, obj. detection and semantic segm.

- In addition to image classification, object detection and semantic segmentation are also used as the downstream tasks to evaluate the quality of the features learned by SSL.
- Usually, ImageNet is used for self-supervised pretext task pre-training by discarding category labels, while the AlexNet is used as the base network and fine-tuned on the three tasks.
- The performance of image classification, object detection, and semantic segmentation tasks are measured on the PASCAL VOC dataset.

61

Image classification, obj. detection and semantic segm.



Method	Pretext Tasks	Classification (%)	Detection (%)	Segmentation (%)
ImageNet Labels [9]	—	79.9	56.8	48.0
Random(Scratch) [9]	—	57.0	44.5	30.1
ContextEncoder [21]	Generation	56.5	44.5	29.7
BiGAN [93]	Generation	60.1	46.9	35.2
ColorfulColorization [20]	Generation	65.9	46.9	35.6
SplitBrain [97]	Generation	67.1	46.7	36.0
RankVideo [110]	Context	63.1	47.2	35.4†
PredictNoise [109]	Context	65.3	49.4	37.1†
JigsawPuzzle [22]	Context	67.6	53.2	37.6
ContextPrediction [45]	Context	65.3	51.1	—
Learning2Count [106]	Context	67.7	51.4	36.6
DeepClustering [31]	**Context**	**73.7**	**55.4**	**45.1**
WatchingVideo [30]	Free Semantic Label	61.0	52.2	—
CrossDomain [53]	Free Semantic Label	68.0	52.6	—
AmbientSound [141]	Cross Modal	61.3	—	—
TiedToEgoMotion [62]	Cross Modal	—	41.7	—
EgoMotion [61]	Cross Modal	54.2	43.9	—

62

Image classification, obj. detection and semantic segm.

- The performance of the self-supervised models on segmentation and detection dataset are very close to that of the supervised method which is trained with ImageNet labels during pre-training.
- Specifically, the margins of the performance differences on the object detection and semantic segmentation tasks are less than 3 percent, which indicate that the learned features by self-supervised learning have a good generalizability.
- In general, context-based methods performance better than other types of methods.

63