VIP

VISUAL INFORMATION PROCESSING BASED ON PAIRWISE LEARNING

許志仲 (Chih-Chung Hsu)
Assistant Professor, cchsu@mail.npust.edu.tw
Department of Management Information Systems,
National Pingtung University of Science and Technology

Outline

- About me
  - Experiences
- DSL: form Anti-GAN to Autonomous Driving applications
  - Overview of Deep Learning
    - Supervised – Unsupervised – Semi-supervised Learning
  - Pairwise Learning based Applications
    - Identity-preserving face hallucination [18-19]
    - Fake face image detection [18-]
    - Risk assessment module for autonomous car [19-]
    - Gastric cancer detection for small-scale M-NBI dataset [19-]
    - Vehicle Re-identification in the wild [19-]

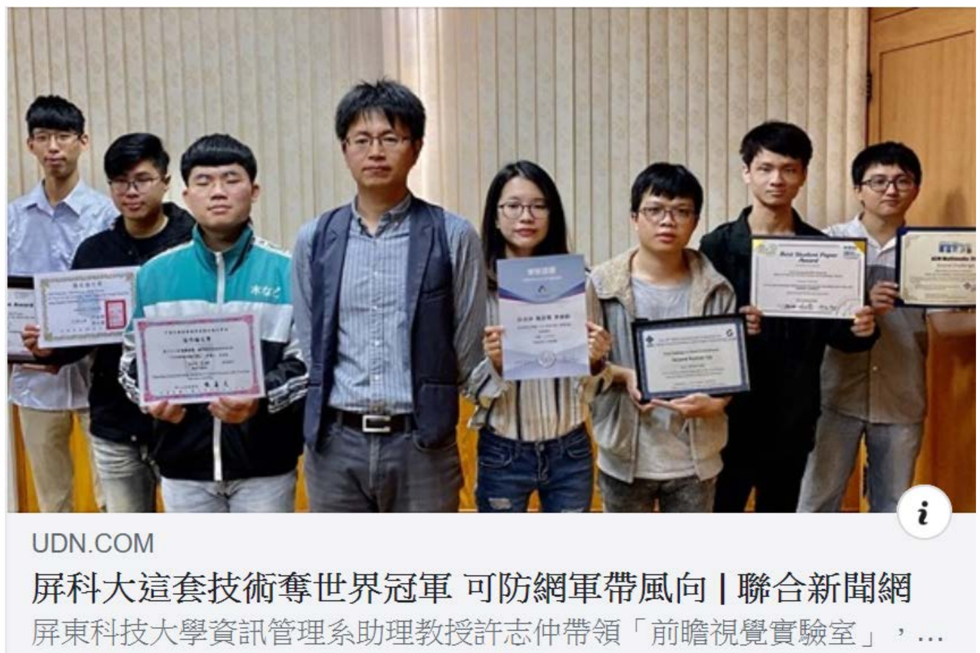
About Me

- Chih-Chung Hsu (許志仲)
  - Assistant Professor, MIS, NPUST

- Selected Experiences
  - IEEE SPS Tainan Chapter Vice Chair, 2020/2-Present
  - Visiting Scholar, NII, Japan, 2017/2.
  - **Co-Founder** & CTO, AI.SKOPY, Incubation Center, HTHU, 2017/10-2018/2
  - **Co-Founder** & Project Director, Eye-Digit. Co., Feb. 2009 – Feb. 2011

Research Summary (UDN News)

- https://udn.com/news/story/7327/4222949?fbclid=IwAR18aZ4Ykj40xrT0WVhL9lwnXXeNVotwcmHV3MnqU0TyRCkvR74RS6eVGVg



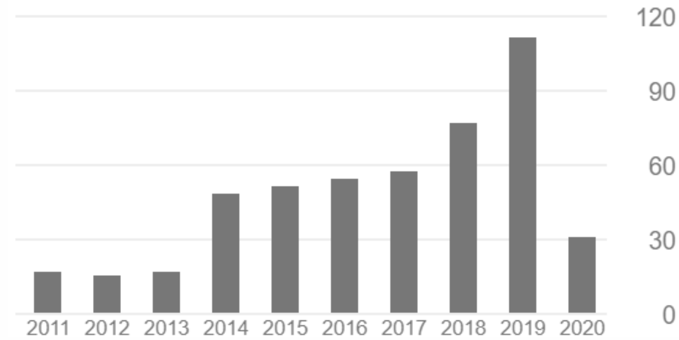
Fake Image Detection (CTEE News)

- https://view.ctee.com.tw/technology/17461.html

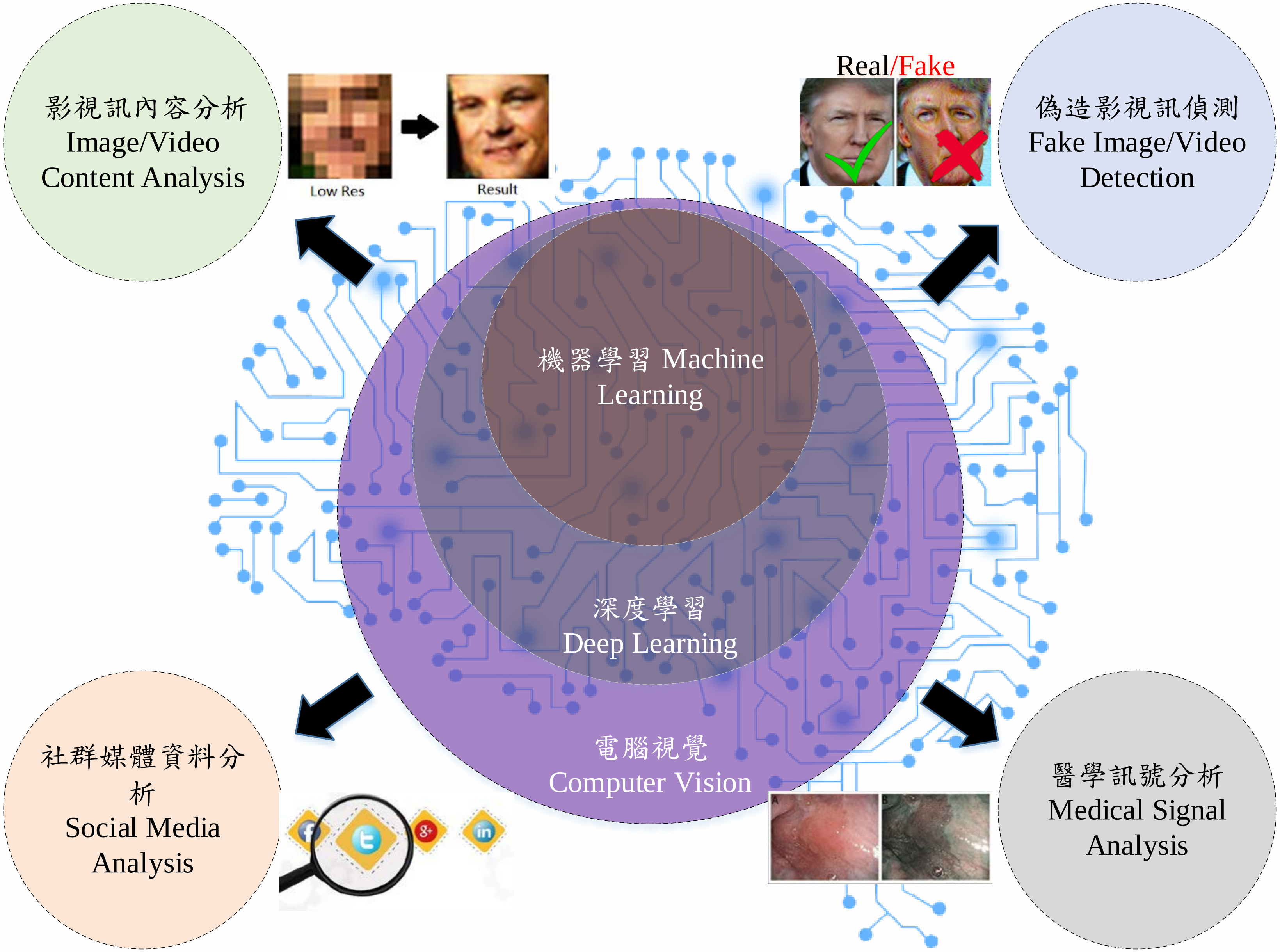
文 / 許志仲、莊易修 國立屏東科技大學資訊管理系助理教授、碩士生

Publication Summary

- Google Scholar
  - # Citations: 506
- Highest one:
  - Video forensic: 181 (since 2008)



- Most influential papers:
  - Fake image detection: 18/year (since 2019)
    - My MOST project
  - Video forensic: 14.6/year


影視訊內容分析
Image/Video
Content Analysis
Low Res
Result
Real/Fake
偽造影視訊偵測
Fake Image/Video
Detection
機器學習 Machine
Learning
深度學習
Deep Learning
電腦視覺
Computer Vision
社群媒體資料分析
Social Media
Analysis
醫學訊號分析
Medical Signal
Analysis

Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

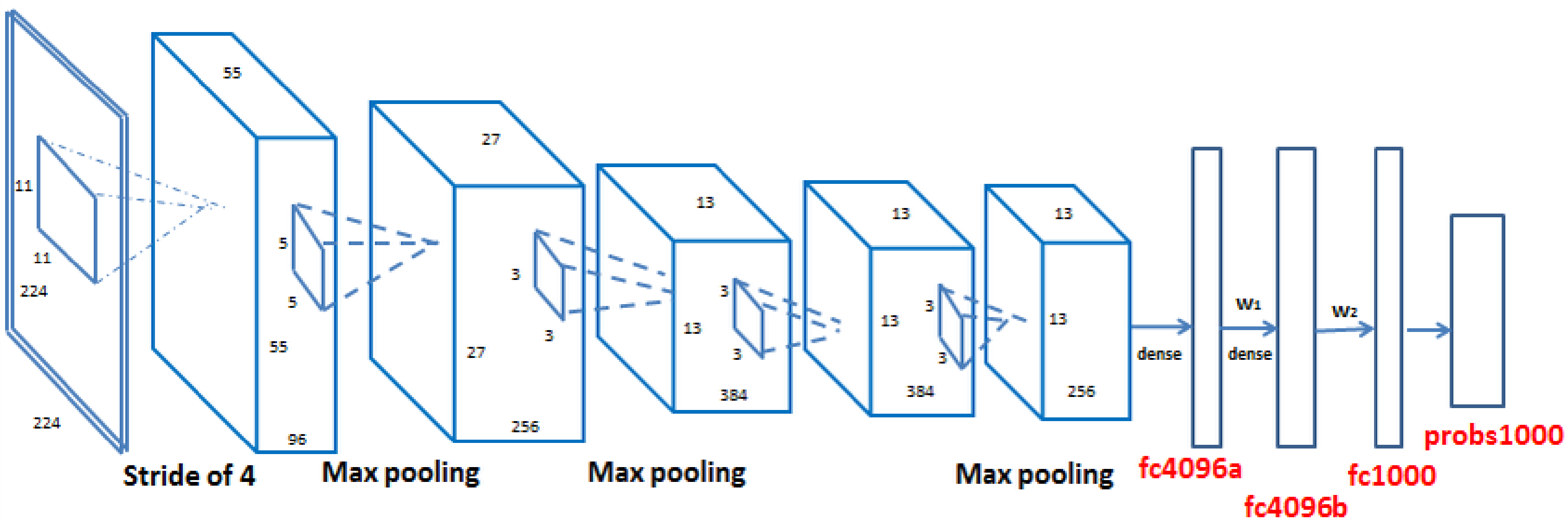
Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

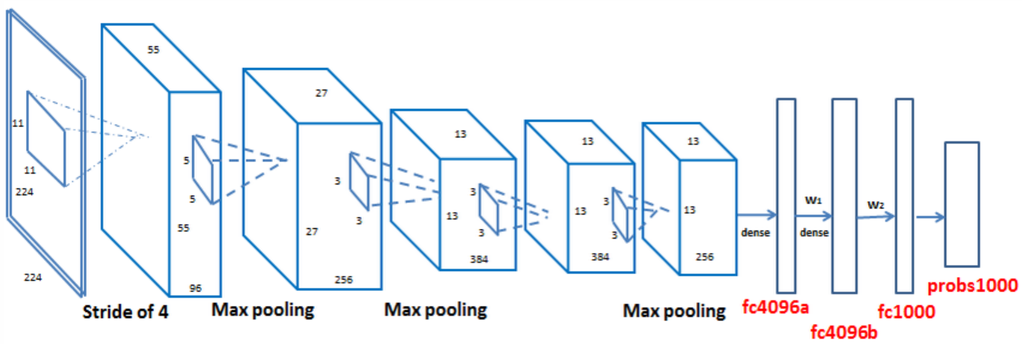
DEEP SUPERVISED LEARNING

AlexNet (2012, Hinton)

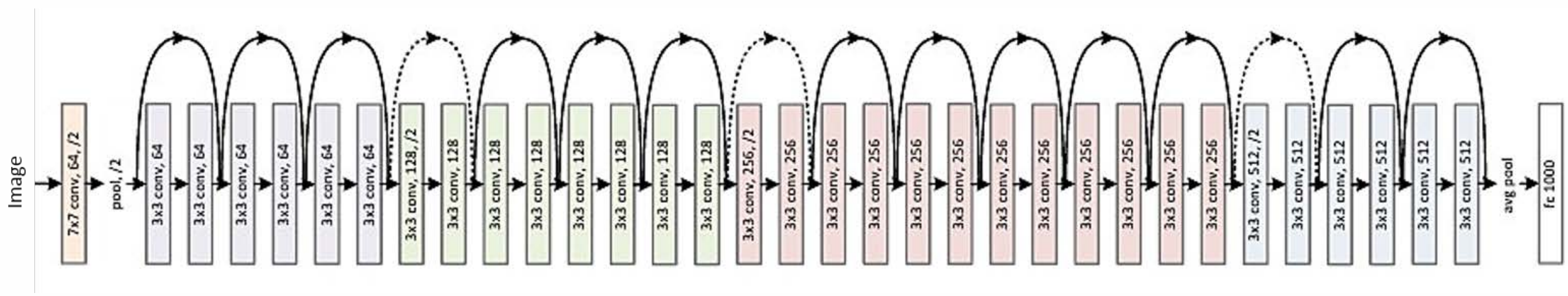
- The winner in ILSVRC Challenge based on Deep Learning in supervised way
- 9-layers
  - 5 convolution and 4 fully-connected layers



Deeper Networks

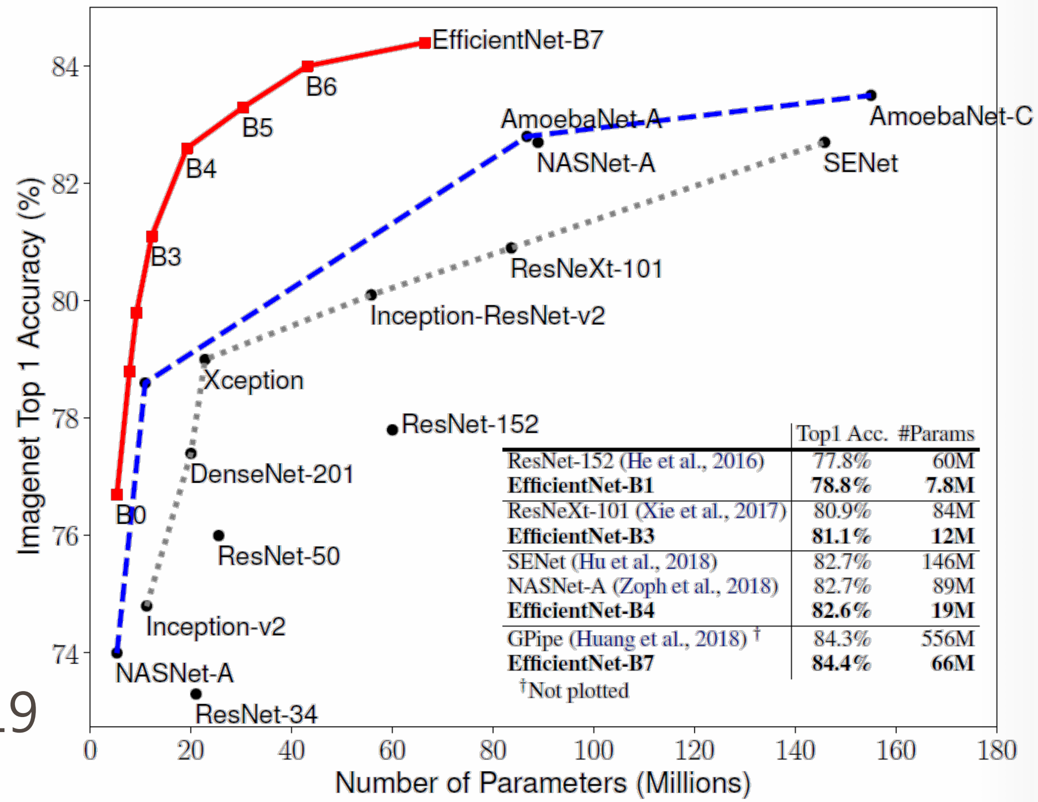


- 2013, AlexNet: 8 layers (9 layers)
- 2016, Residual Net / DenseNet: up to 152 layers…
- 2017, Stochastic depth Net: up to 1000 layers…



State-of-the-Art CNNs

- We called those CNNs trained in supervision way are “backbone ” or “baseline” nets
- SOTA now
  - High-performance
    - ResNet
    - Wide-ResNet
    - ResNeXt
    - Inception v3
    - DenseNet
  - High-efficiency
    - MobileNet v3
    - EfficientNet
- Anti-aliasing CNNs ICML19



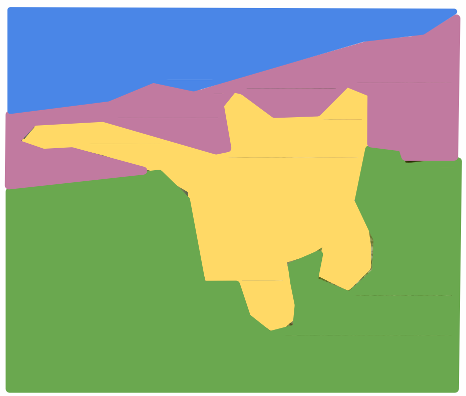
Computer Vision Applications

Classification

CAT

No spatial extent

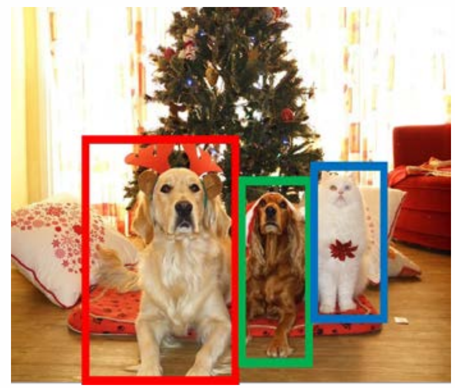
Semantic Segmentation



GRASS, CAT, TREE, SKY

No objects, just pixels

Object Detection



DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

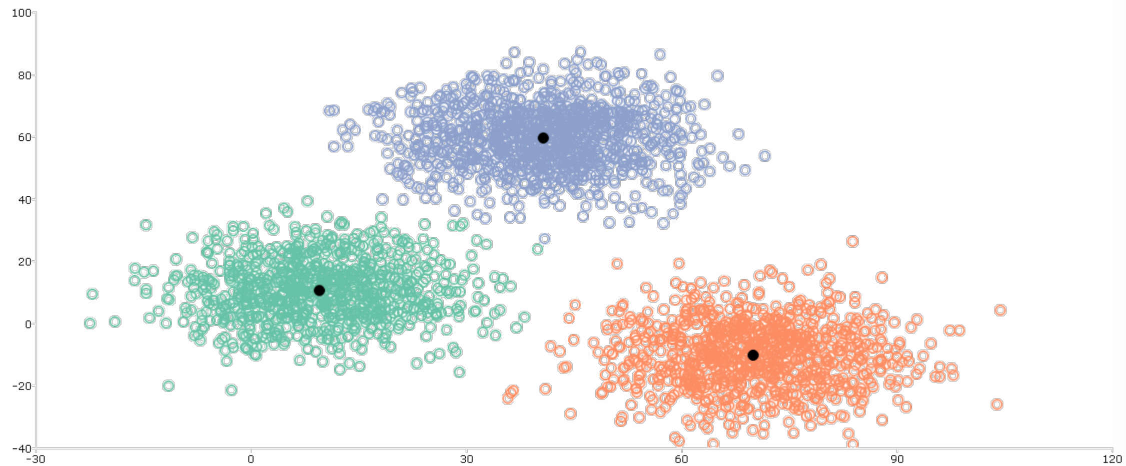
Multiple Object

Slide credit: CS231n, Stanford

DEEP UNSUPERVISED LEARNING

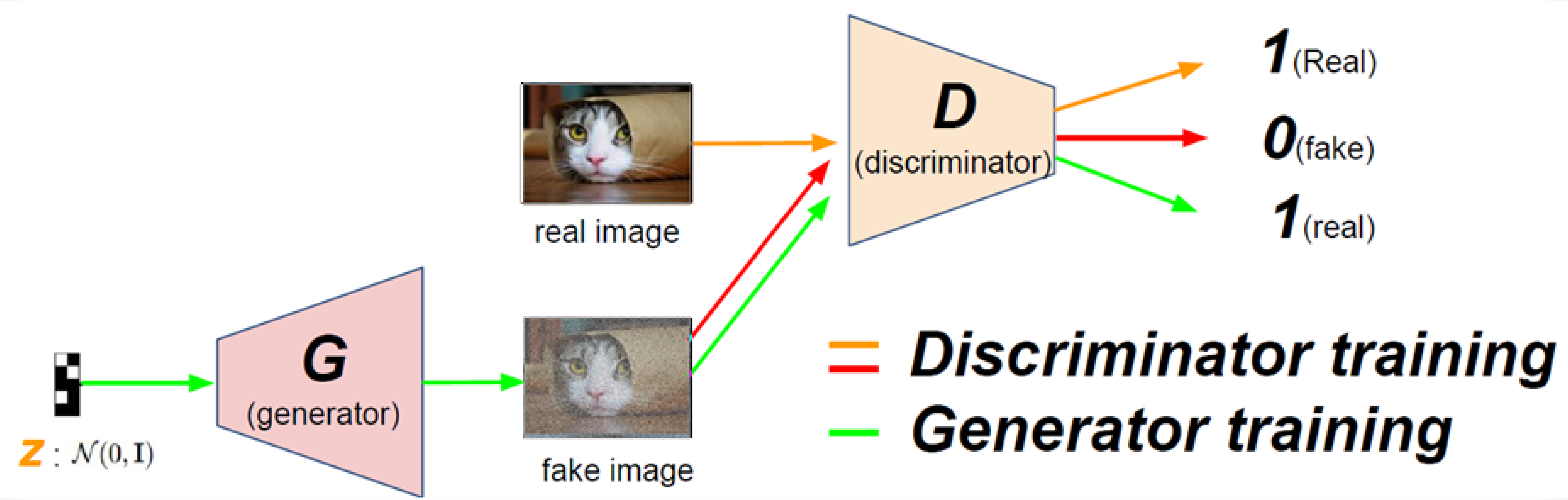
Unsupervised Learning

- Feature representation
  - Dimensionality reduction
    - High-dimensional data ➔ Low-dimensional one
- Generative model
  - Low-dimensional data ➔ High-dimensional one
- Clustering
  - Data analysis



Unsupervised Deep Learning

- How to generate an image with good quality?
  - Generative adversarial network (GAN)



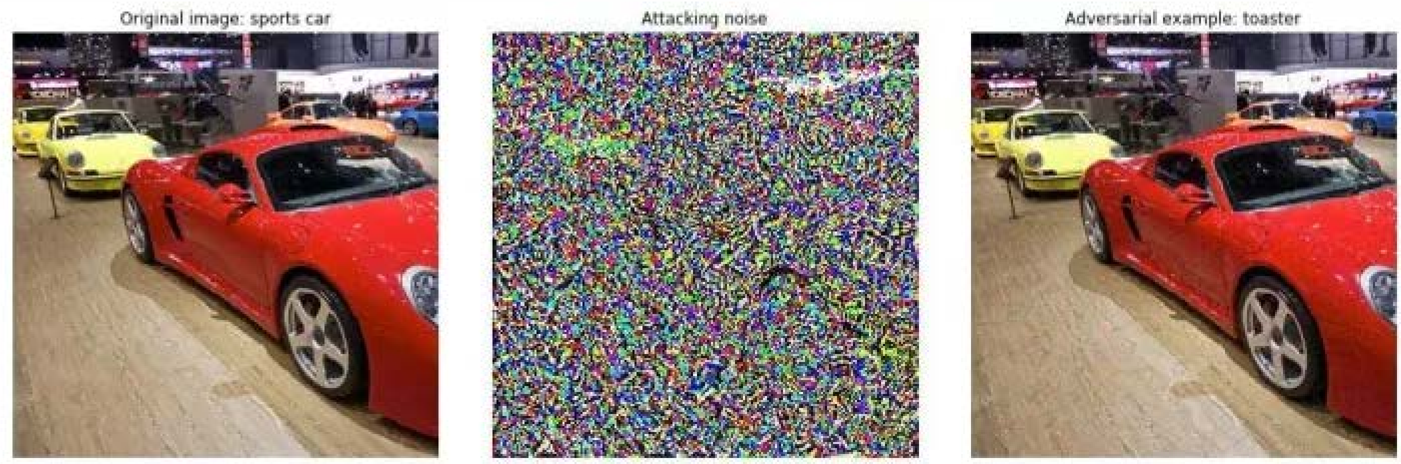
Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

換臉 (Spiderman – 2016)

https://youtu.be/kxgqt6-0dck

Rethinking GANs

- Is possible to fool a DNN by adding specified noises?
  - Adversarial attack



(a) Car (b) Noise map (c) Toaster

SEMI-SUPERVISED LEARNING

Incorporating partial label information

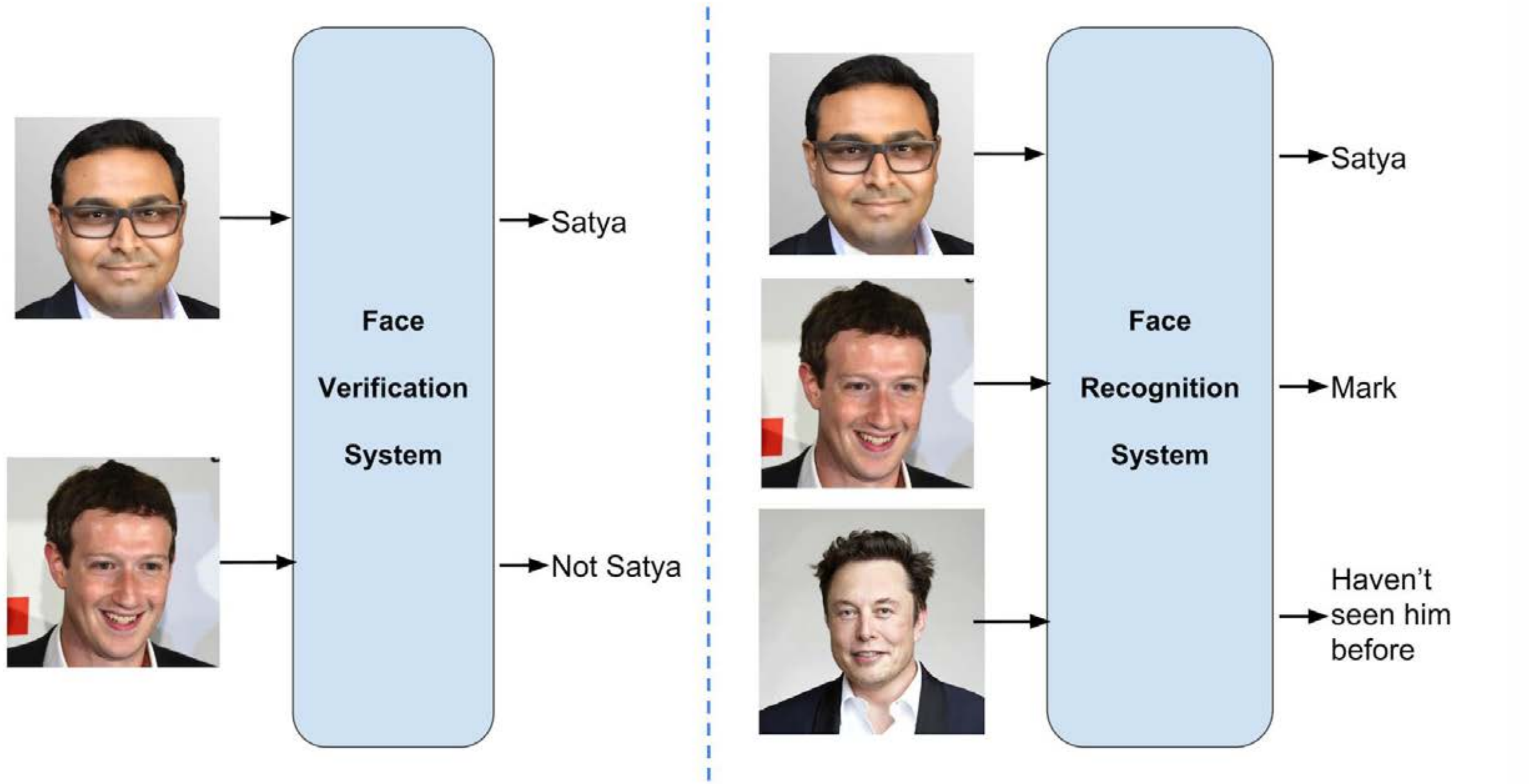
Deep Semi-Supervised Learning (DSL)

- Take some advantages form supervised/unsupervised learning
  - Problem: How?
- Definition of DSL
  - Given a dataset with partial label information
    - Partial data have labels (Few-shot learning)
      - Usually EM can be used to solve this problem
      - Initial model can be learning based on labeled data (Transfer learning)
      - Get pseudo labels of unlabeled data using the model (MixMatch, 19’)
      - Re-training model and repeat…
      - Others: Label-propagation… (Siamese networks)
    - Partial label information only (i.e., same/different identity)
      - Data can be augmented
      - Siamese Network [LeCun 05]

Siamese Network

- It is easy to learn from the limited samples
  - Real-world applications
    - Data may have few labels…
    - E.g. 1000 classes, 5 images/class = 50,000 samples
- Siamese Network
  - Pairwise Learning
  - Make data "Pairwise"
    - Same identity of a pair: y=1
    - Different identities of a pair: y=0
    - 50,000 samples ➔ C(1000,2)*5 = 2,497,500 pairs
  - Usually used in "face verification" or person re-identifications

Face Verification versus Face Recognition



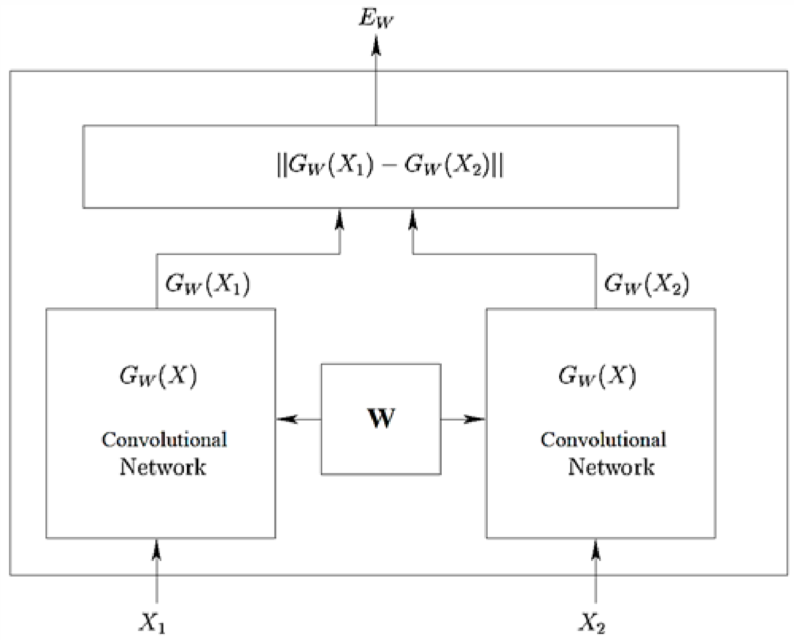
Siamese Network

- Key to face verification
  - Discriminative feature representation
    - A pair with the same identity
      - Features should be similar to each other
    - A pair with the different identities
      - Features should be different from each other

- Applications
  - Few-shot learning (learn features from the limited training samples)
    - Based on pairwise learning or the loss functions from rank/metric learning

Siamese Network (cont.)

- Siamese Network Architecture
  - Learning to capture the discriminative feature
  - Simply minimizing the distance between two samples with the same identity



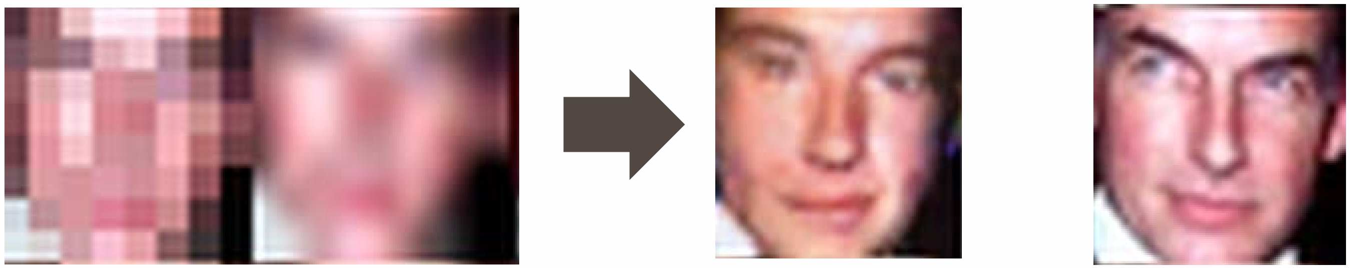
Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

IDENTITY-PRESERVING FACE HALLUCINATION

ICIP 18, IEEE Transactions on Image Processing (TIP), Dec. 2019.
Contribute to my MOST Project

Traditional Face Hallucination

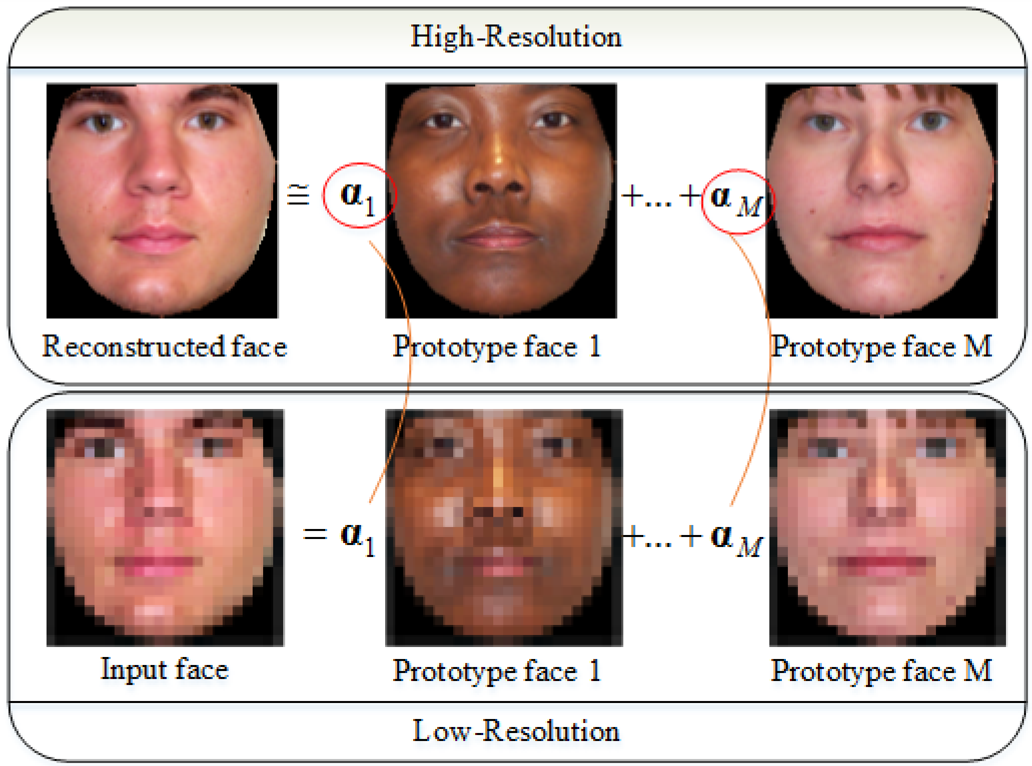


LR Bicubic SR HR

Amazing but identity unrecognizable!

We achieve

Face Hallucination



$$\mathbf{I} \cong \boxed{\mathbf{P}} \cdot \boldsymbol{\alpha} = \mathbf{R}$$

$$\boldsymbol{\alpha}^* = ((\mathbf{P}_L)^T \cdot \mathbf{P}_L)^{-1} \cdot (\mathbf{P}_L)^T \cdot \mathbf{I}_L$$

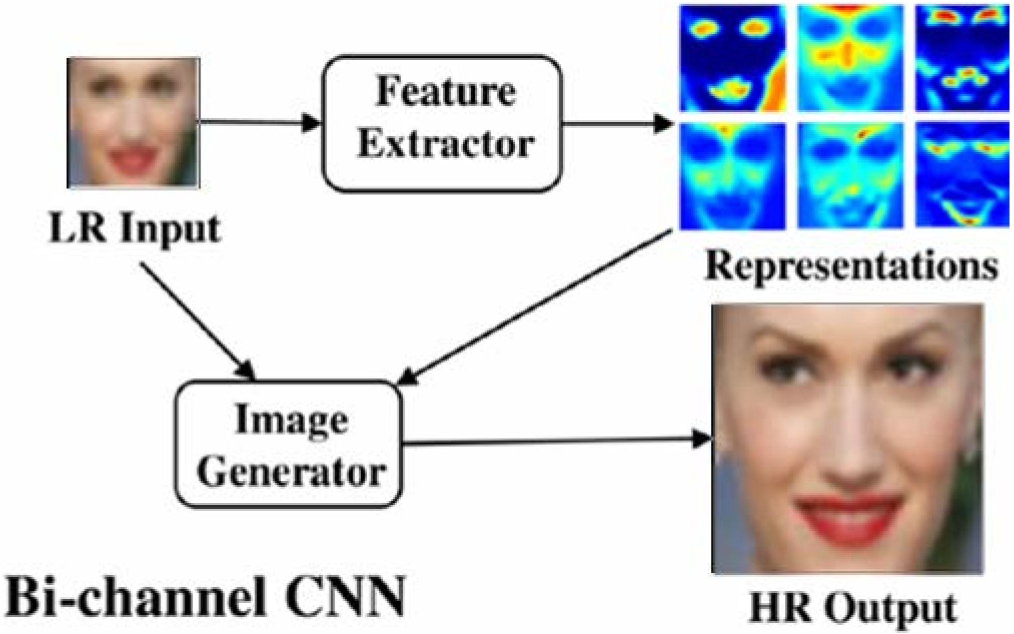
Dictionary

Learning to Hallucinating Face

- Traditional approach
  - Dictionary learning by PCA, NMF, ONMF,…etc
- Deep learning-based approach
  - End-to-end architecture
    - Input low-resolution face image, out high-resolution face image directly.
- Deep neural network has different structures
  - CNN-based (Convolutional neural network)
    - Upsampling layer upscales input signal
  - GAN-based (Generative adversarial network)
    - High quality result
    - May result in identity-unrecognizable

CNN-based Approach (AAAI' 15)

- Using CNN to learn the dictionary and its coefficients

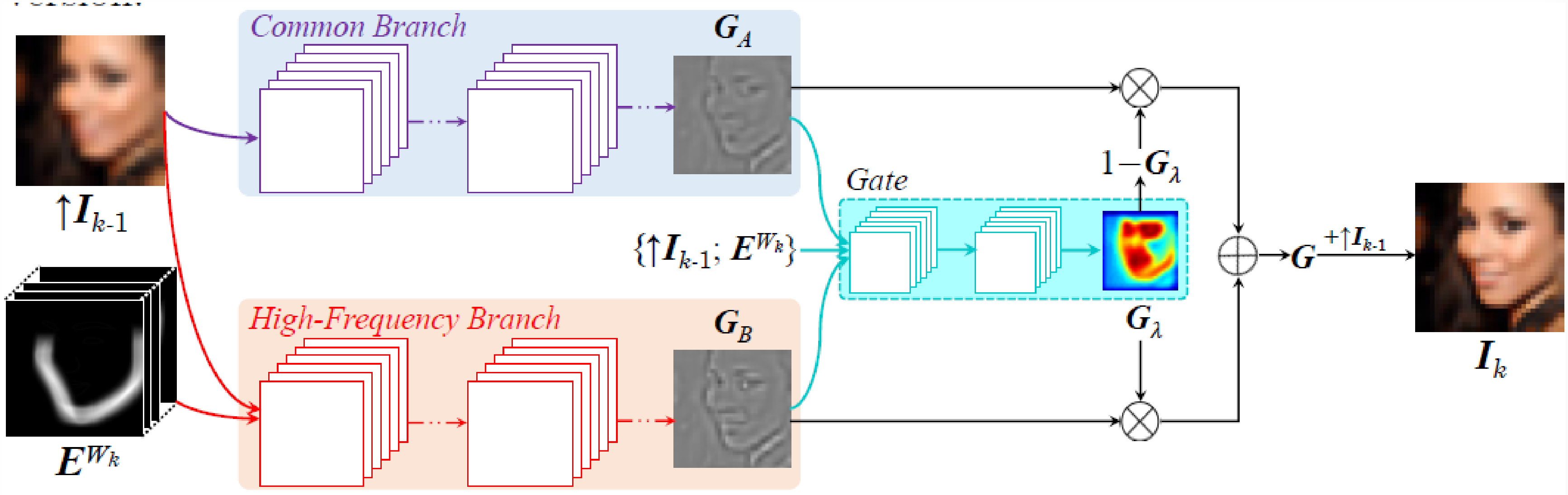


Zhou, Erjin, et al. "Learning Face Hallucination in the Wild." *AAAI*. 2015.

CNN-based Approach (AAAI' 15)

- Pros
  - First approach based on deep neural network (DNN)
  - Alignment is unnecessary
  - State-of-the-art result (2015)
- Cons
  - The visual quality of reconstructed face image will be poor when
    - Extreme low-resolution
      - i.e. 8x8
    - Identity-unrecognizable

Cascaded CNN Approach (ECCV'16)



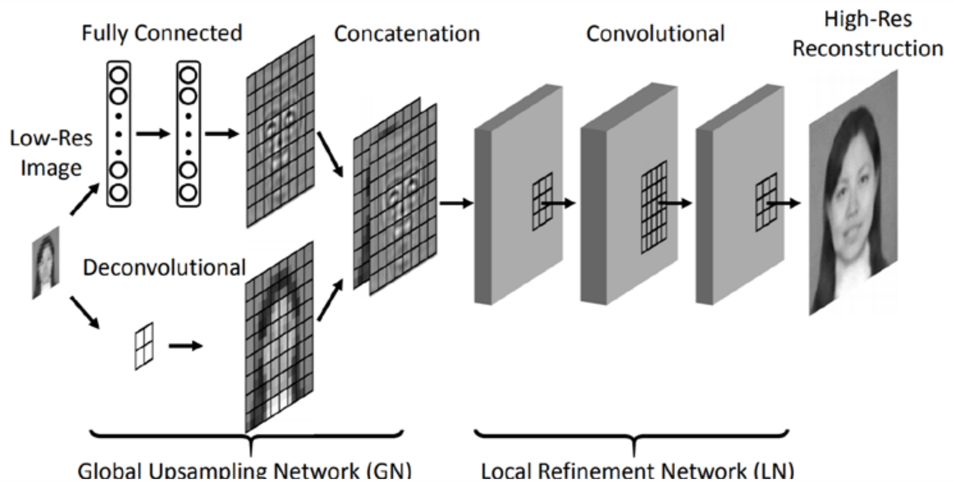
- Cascaded multiple CNN to enhance visual quality
- Gate network can be used to fusion of two nets

Zhu, Shizhan, et al. "Deep cascaded bi-network for face hallucination." *European Conference on Computer Vision*. Springer International Publishing, 2016.

Cascaded CNN Approach (ECCV’ 16)

- Pros
  - The best performance so far
  - Alignment-free
  - More realistic
- Cons
  - It is very hard to train
    - Released code has no training codes
  - A lot of parameters need to be tuned manually
  - Extreme low-resolution inputs
    - Cannot obtain promising results

GAN (Generative Adversarial Net) for Face Hallucination

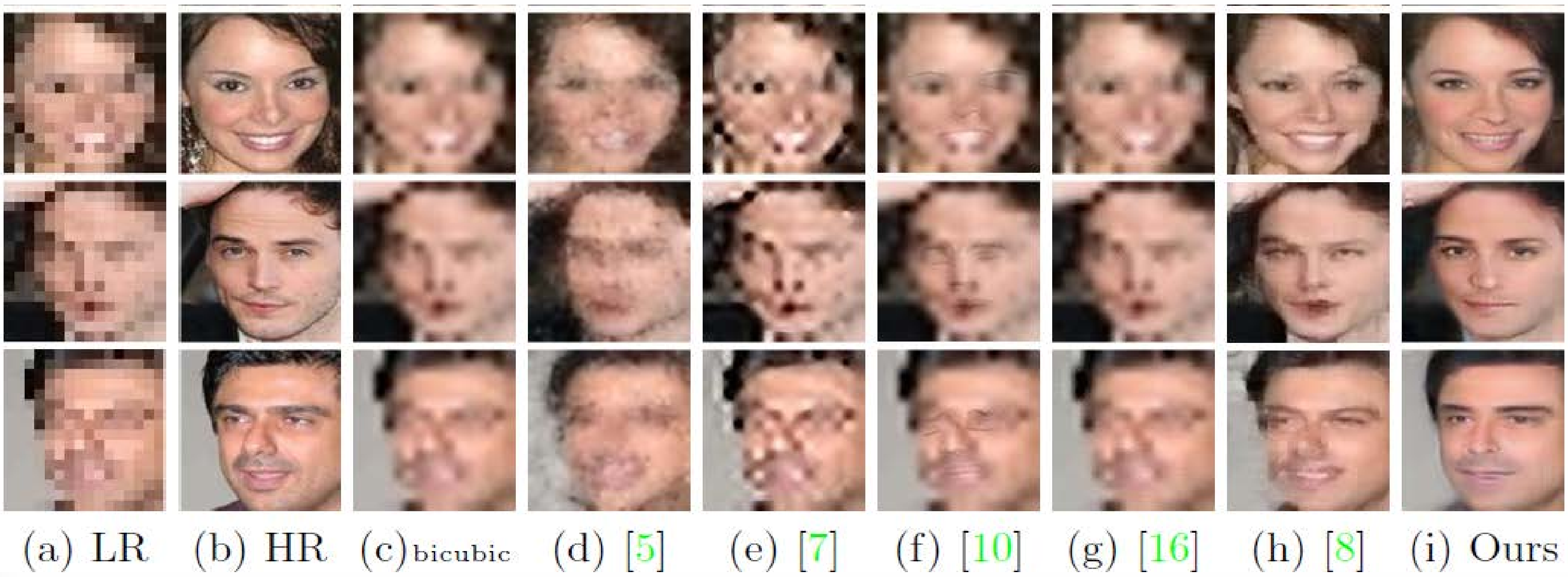


- Use discriminator to refine the upsampling network
  - Dissimilar to the ground truth

Tuzel, Oncel, Yuichi Taguchi, and John R. Hershey. "Global-Local Face Upsampling Network." *arXiv preprint arXiv:1603.07235* (2016). [no code]

GAN for Face Hallucination (II)

- Discriminator is used to judge the visual quality



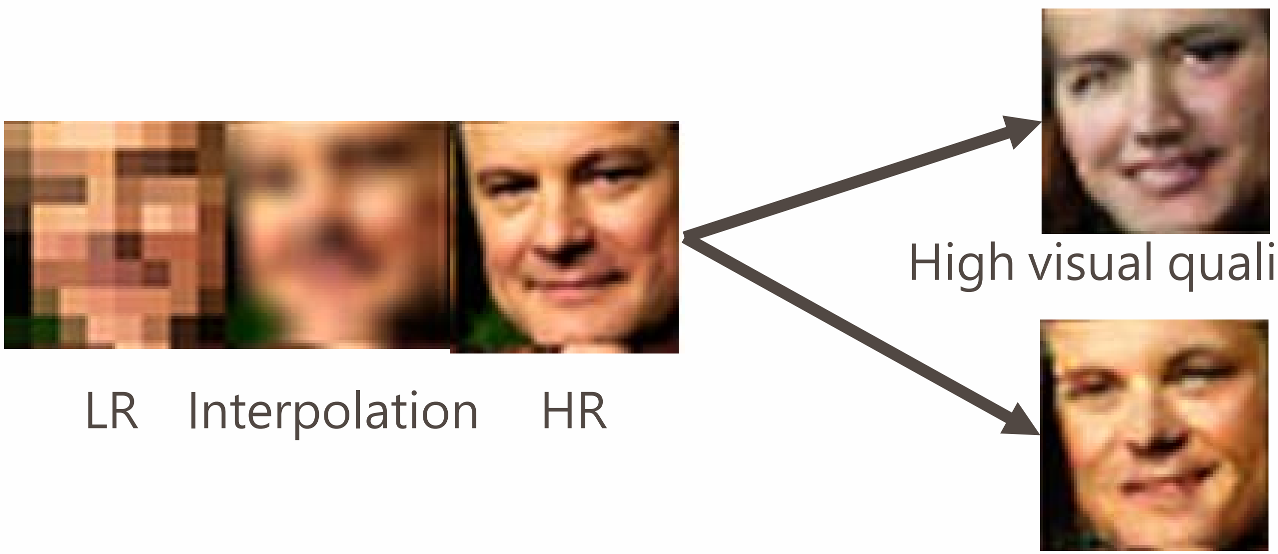
Yu, Xin, and Fatih Porikli. "Ultra-resolving face images by discriminative generative networks." *ECCV*, 2016. [no code]

GAN-based Face Hallucination

- Pros:
  - High visual quality of the reconstructed image
- Cons:
  - May be identity-unrecognizable

Our Goal

- High visual quality reconstruction
  - Even in extreme low-resolution inputs
- Identity-recognizable reconstruction
  - As similar to the ground truth as possible

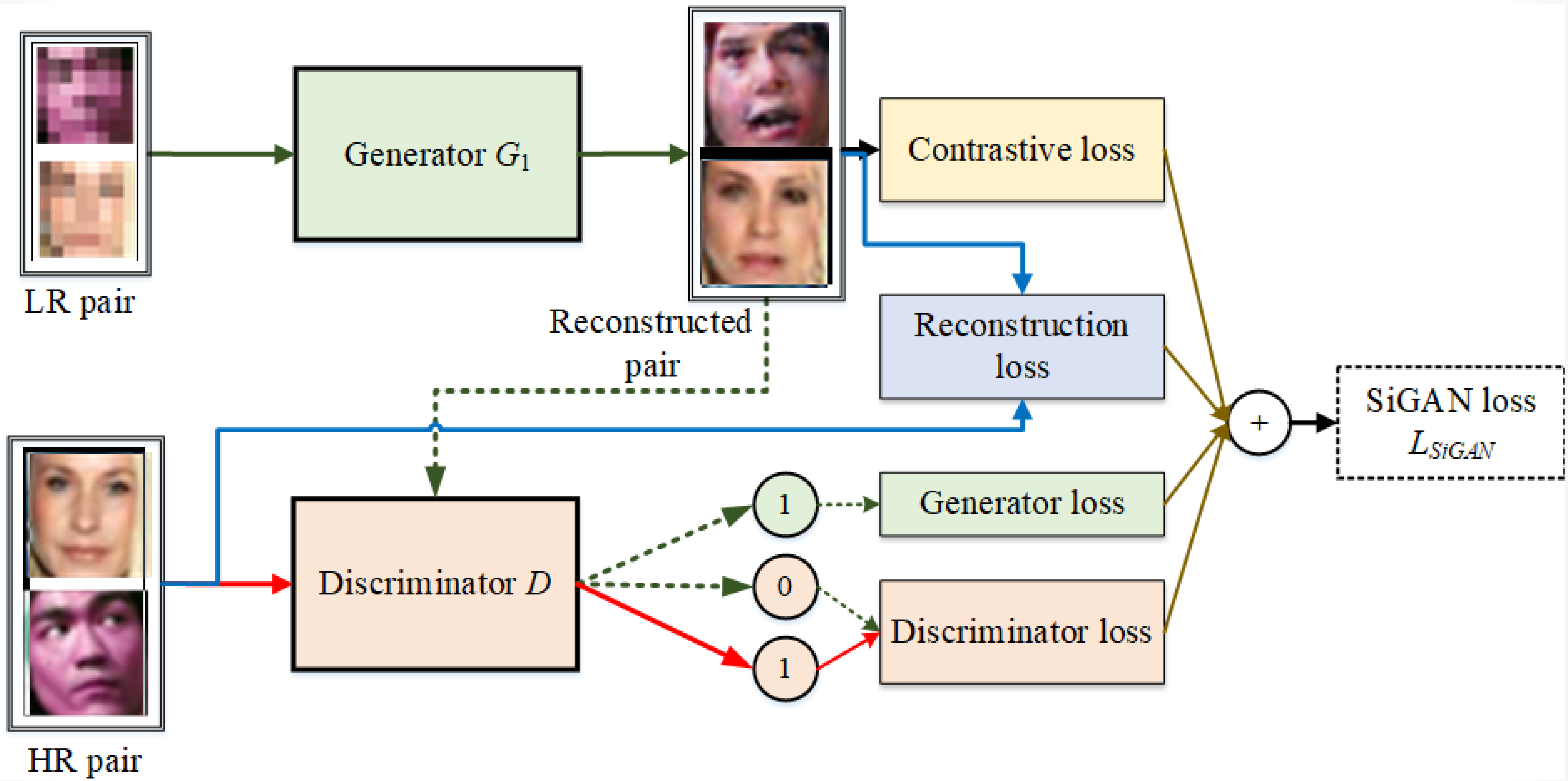


Identity-recognizable & high visual quality

Our Solution

- Key idea
  - Label embedding
    - Use the label information to fine-tune the generator
    - Identity-recognizable reconstruction
  - We propose "Siamese GAN" (SiGAN)
    - Label information will guide the "generator" how to obtain both high-visual quality and identity-recognizable result
    - Partial label information needs only

The Proposed SiGAN

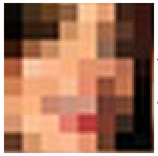
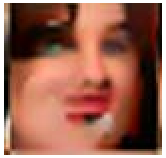


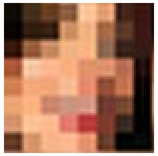
The Loss Function of The Proposed SiGAN

- Loss function for our generator

$$\min_G \max_D V(D,G) = E_D\left[\log D(\mathbf{x}_1^{HR})\right]$$
$$+E_G\left[\log\left(1 - D(G(\mathbf{x}_1^{LR}))\right)\right] + E_C\left[G(\mathbf{x}_1^{LR}), G(\mathbf{x}_2^{LR})\right],$$

  - subject to $\|y^{HR} - y^{SR}\|_1 < \epsilon$
  - SR result: G($\mathbf{x}^{LR}$)
  - E_C represents contrastive loss

D [G() = ] = 0

D [G() = ] = 1

Contrastive Loss for SiGAN

- If we directly minimize Ew(X1, X2)
  - The energy and the loss can be made zero by simply making Gw(X1) a constant function
  - We don’ t want to see that
- By adding a contrastive term
  - The loss function can be

CNN’ s parameters

The same or not (0/1)

Partial loss function for a genuine pair

$$L(\mathrm{W}) = \sum_{i=1}^{P} L(\boxed{W}, \boxed{(Y,} \boldsymbol{x}_1, \boldsymbol{x}_2)^i)$$

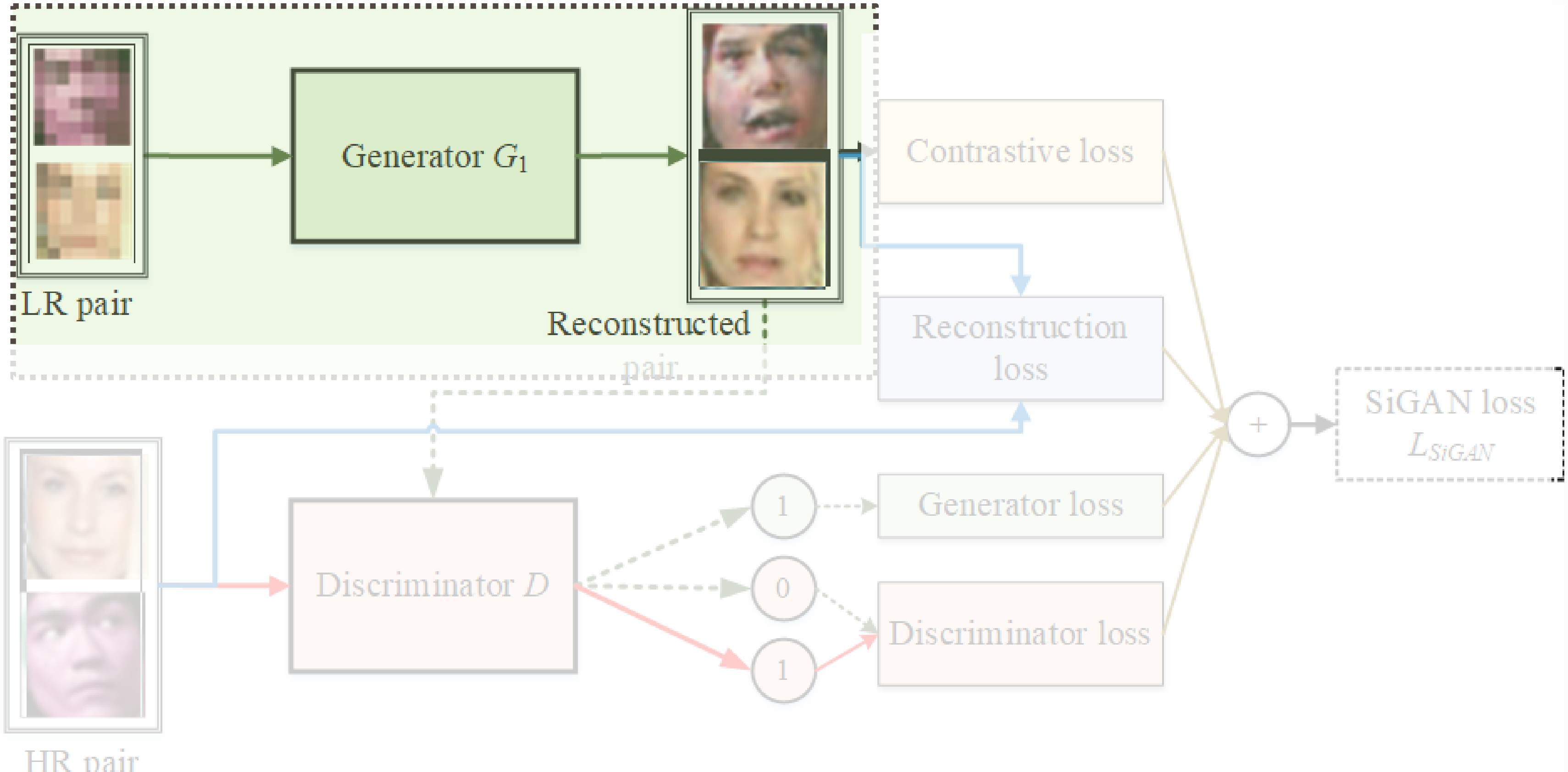
$$L(W, (Y, \boldsymbol{x}_1, \boldsymbol{x}_2)^i) = y \boxed{L_G} \big(E_w(\boldsymbol{x}_1, \boldsymbol{x}_2)\big) + (1-y) \boxed{L_I} (E_w(\boldsymbol{x}_1, \boldsymbol{x}_2))$$

$$L_{\mathrm{G}} = \frac{1}{2}(E_w)^2$$

$$L_{\mathrm{I}} = \frac{1}{2}[\max(0, margin - E_w)]^2$$

Partial loss function for an impostor pair

Test Stage of The Proposed SiGAN



A simple forward process

Experiment Settings

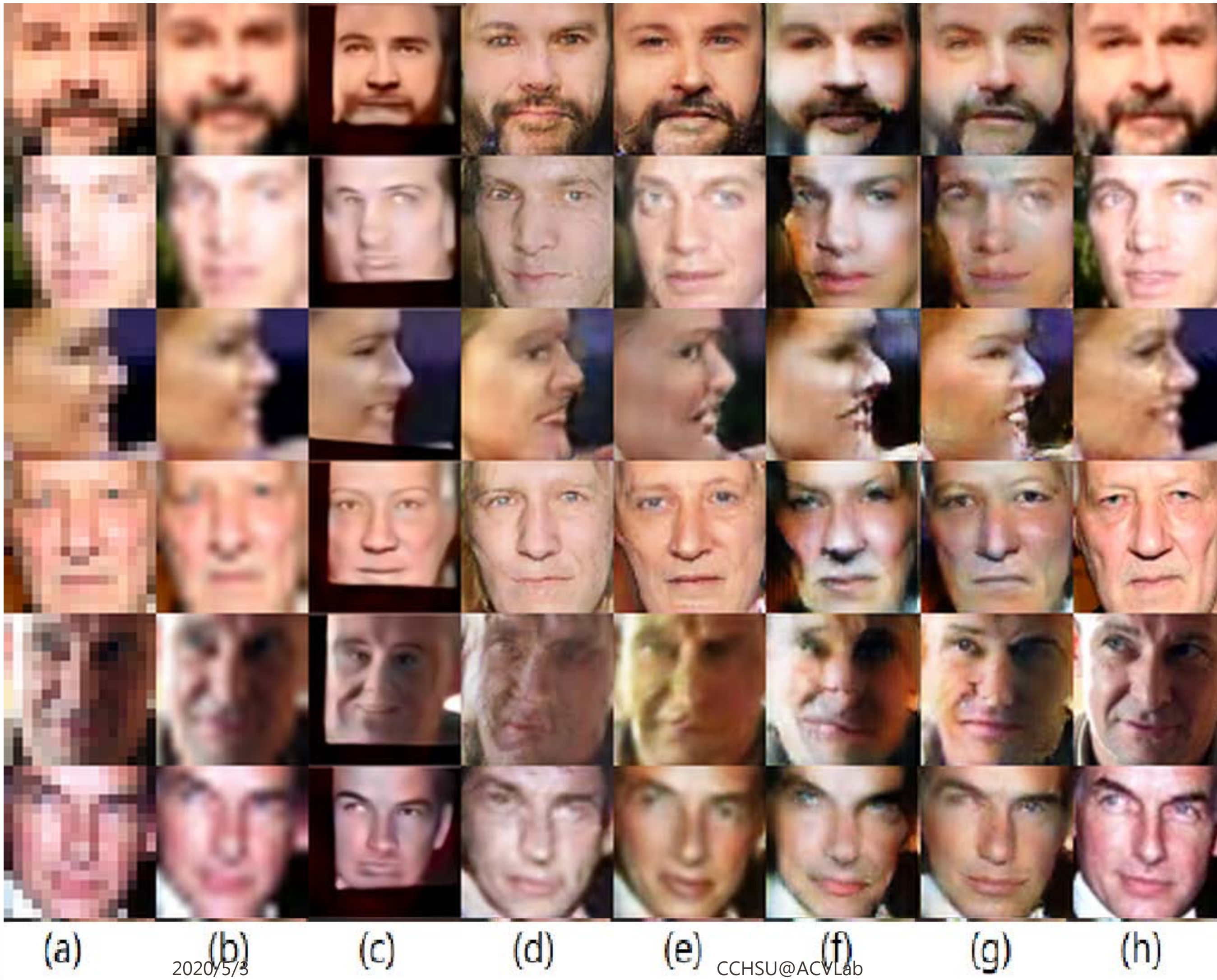
- LR: 8x8
- HR: 32x32 (4x upscaling factor)
- #Identities of training set: 10,575
- #Training images: 491,131
- #Test images: 3,283
- Face recognition engine: FACENET (State-of-the-art)

Subjective Result (8x8➔32x32)

- Face hallucination: Identity-recognizable reconstruction

(a) LR face image
(b) Bicubic interpolation
(c) ECCVCNN
(d) GAN
(e) UR-DGN
(f) Pixel-Recurrent (Google)
(g) Ours w/o label
(h) Ours
(i) Original HR face image

Subjective Result (16x16→64x64)



(a) LR face image
(b) Bicubic interpolation
(c) ECCVCNN
(d) GAN
(e) UR-DGN
(f) Pixel-Recurrent (Google)
(g) Ours
(h) Original HR face image

Objective Results

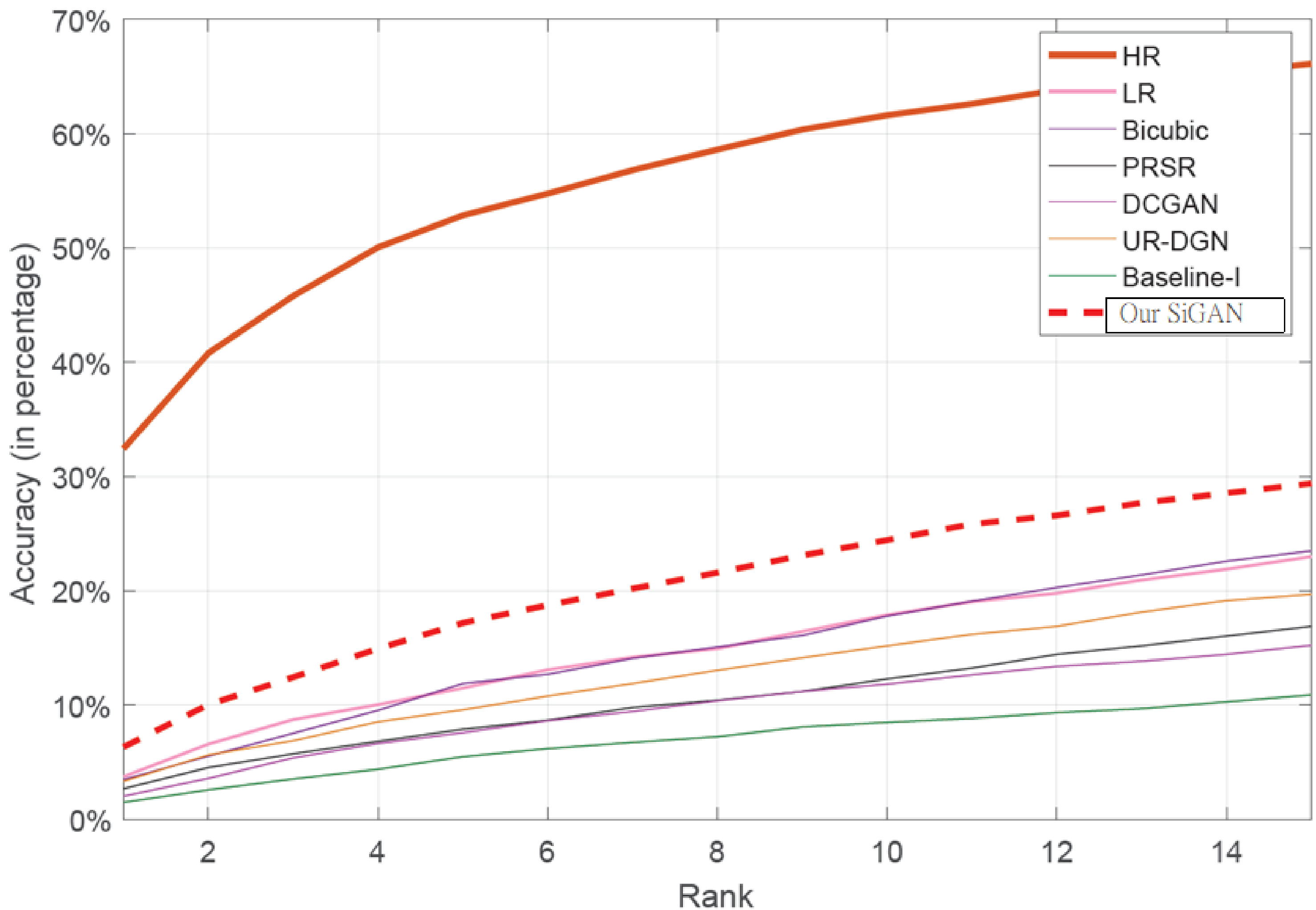
Method	Top-1	Top-5	Top-10
HR	32.4%	52.8%	61.6%
LR	3.7%	11.5%	17.9%
Bicubic	3.5%	11.9%	17.8%
CBN [22]	2.2%	7.8%	12.7%
UR-DGN [21]	3.4%	9.6%	15.2%
DCGAN [15]	2.0%	7.6%	11.8%
PRSR [5]	2.7%	7.9%	12.3%
Ours	6.4%	17.2%	24.5%

Face recognition rate comparison
LR=8x8
HR=32x32

Face recognition rate comparison
LR=16x16
HR=64x64

Method	Top-1	Top-5	Top-10
HR	36.8%	55.9%	63.8%
LR	12.4%	27.4%	37.1%
Bicubic	11.6%	27.5%	37.6%
CBN [8]	3.4%	9.9%	15.4%
UR-DGN [7]	12.2%	29.0%	38.7%
DCGAN [5]	9.3%	24.9%	33.9%
LeGAN (proposed)	17.0%	36.3%	46.4%

Objective Result (8x8)



Summary of Our SiGAN

- Contributions
  - Label information is embedded in the generator of GAN
    - A Guider for the generator
  - High visual quality and identity-recognizable reconstruction
  - Faster hallucination process

LR	Traditional SR	Ours	Ground Truth

Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

FAKE IMAGE DETECTION: ANTI-GAN

IS3C 2018, ICIP 2019*, Journal of Applied Sciences (SCI, Q1)
ICIP Best Student Paper Award (2071 submissions)
Contribute to my MOST project
High impact papers

Detecting the Fake Images

- The related techniques to detect the fake images
  - Intrinsic feature based approach
    - Image forensic
    - Image forgery detection
  - Extrinsic feature based approach: Watermarking
- Intrinsic feature based approach is relatively practical
  - However, such generated images didn't have such intrinsic features
    - Image is generated directly from noise
      - No source

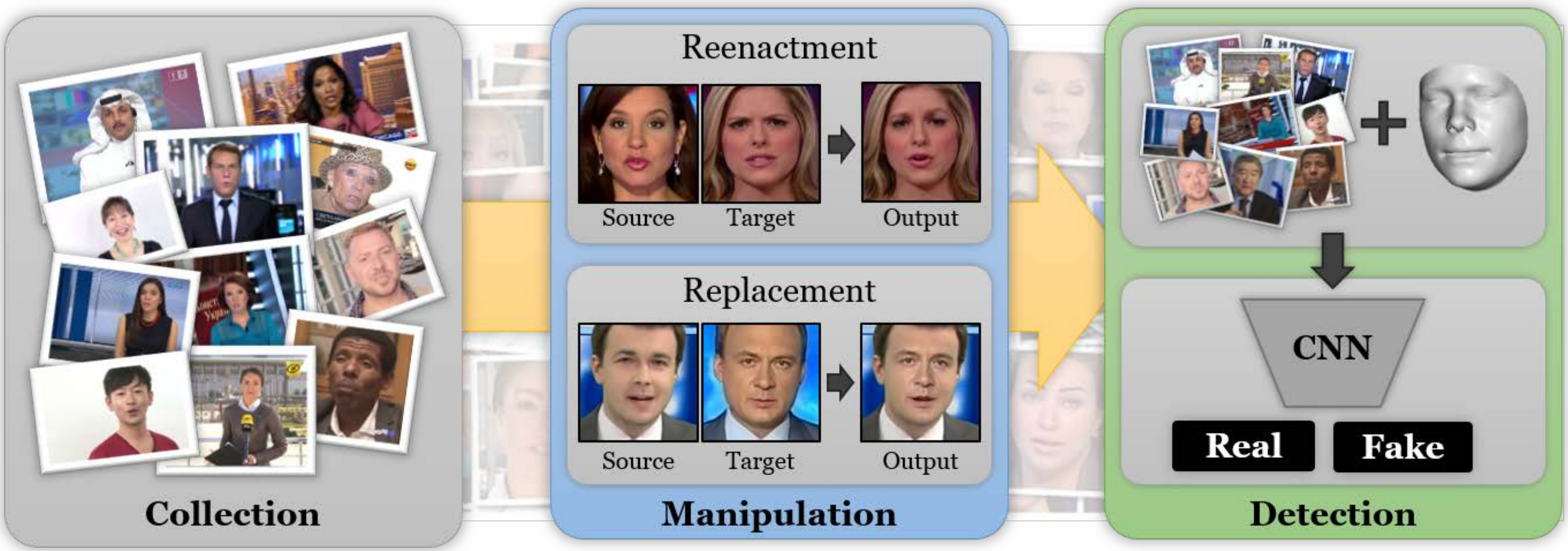
Problems Caused by Fake Images

- Improper use of such fake multimedia will lead to a serious consequence

  - Police purpose, on purpose misleading, or business use

FaceForensic++

- Google provides a large-scale fake image dataset (2019/9)
  - Our initial work was published in 2018/10
- DeepFake Challenge (hosted by Kaggle since 2020/2)
  - AWS, Facebook, Microsoft



An Example of Traditional Image Forensic

(a) Original Image 1 (b) Texture replaced

An Example of Traditional Image Forensic

(a) Fake Image 1 (b) Fake Image 2

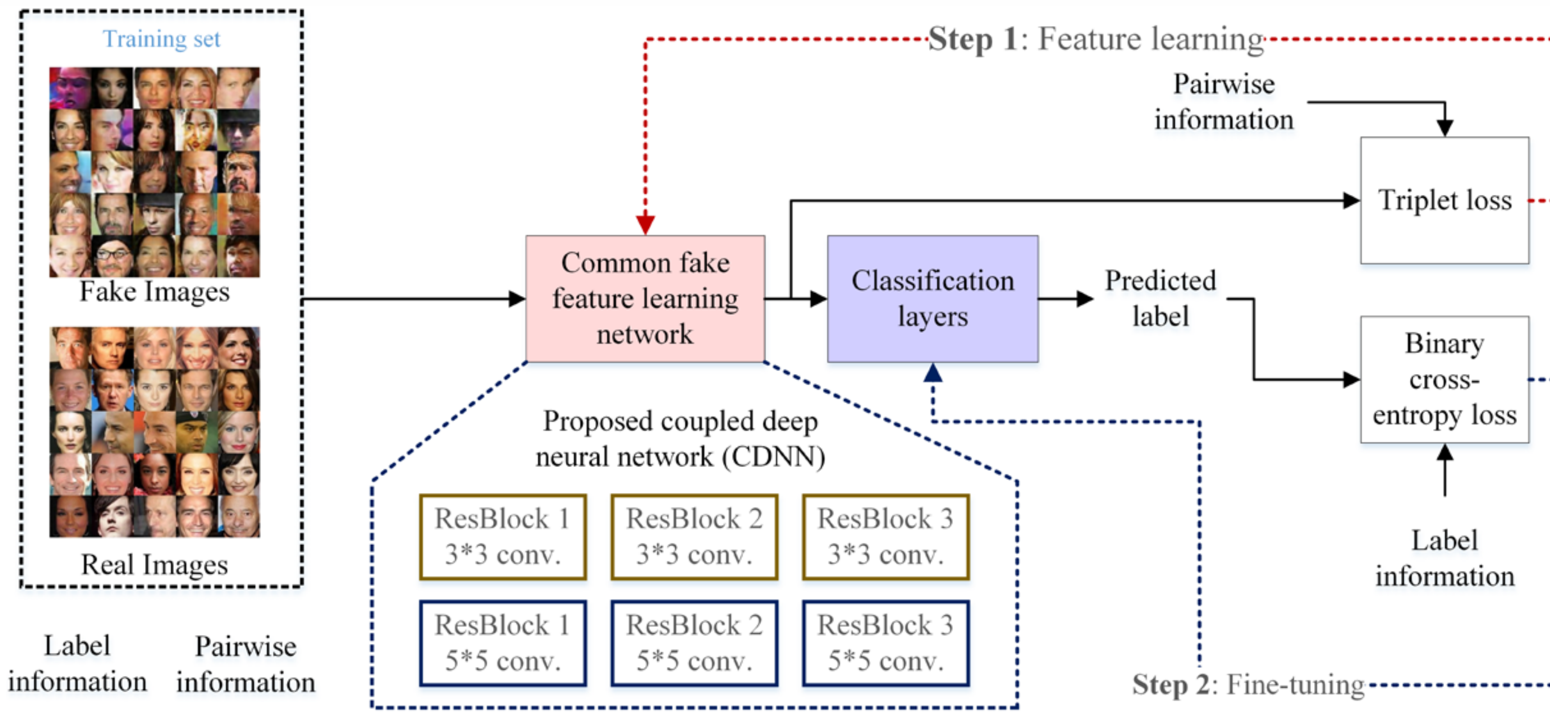
How to effectively detect such fake images remains big problem!!

We propose a novel framework to effectively address this issue!!

Fake Image Detection

- Directly learning a classifier in supervised learning manner may be ineffective.
  - It is hard to collect all GANs to learn
  - The generator can be improved
    - The fake image detector should be improved as well
    - It is too impractical
- Instead of supervised learning, we adopt pairwise learning to effectively capture the common features across different GANs
  - Pairwise learning (PL)
  - Two-step learning policy
    - Called deep forgery detector (DeepFD)

The Proposed Framework



PL1: Contrastive Loss

- Minimizing the feature distance between the paired inputs if they are all fake or real.

$$\mathrm{E}_W(\mathbf{x}_1, \mathbf{x}_2) = \|\mathrm{D}_1(\mathbf{x}_1) - \mathrm{D}_1(\mathbf{x}_2)\|,$$

  - Where D indicates feature representation of JDF of an image
- The contrastive loss function of the proposed JDF will be:

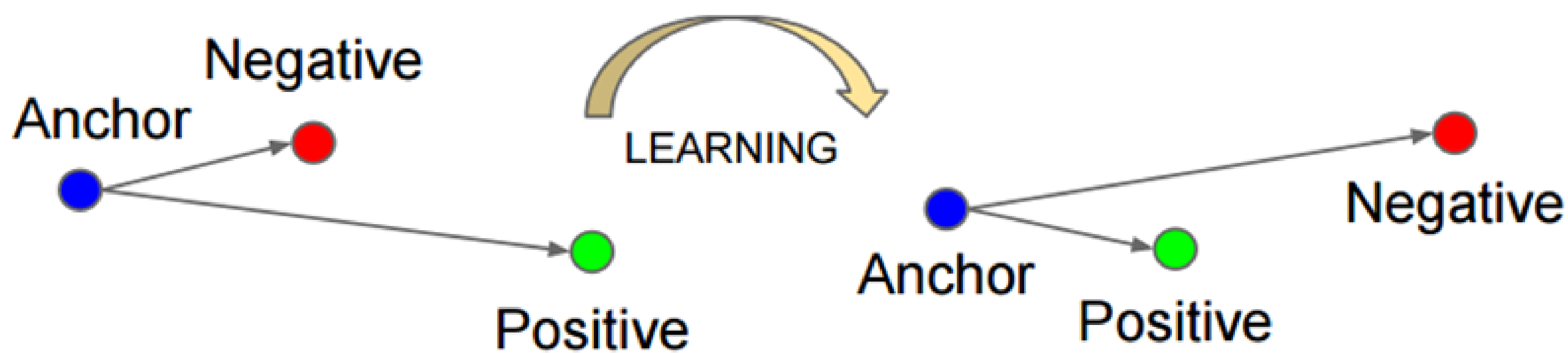
$$L(W, (P, \mathbf{x}_1, \mathbf{x}_2)) = \frac{1}{2}\left(p_{ij}(\mathrm{E}_W)^2 + (1 - p_{ij})(\max(0, m - \mathrm{E}_W)^2\right),$$

  - where p_{ij} indicates genuine ($p_{ij} = 1$) and impostor ($p_{ij} = 0$) pairs

PL2: Triplet Loss

- Calculate the distance between anchor and positive/negative samples

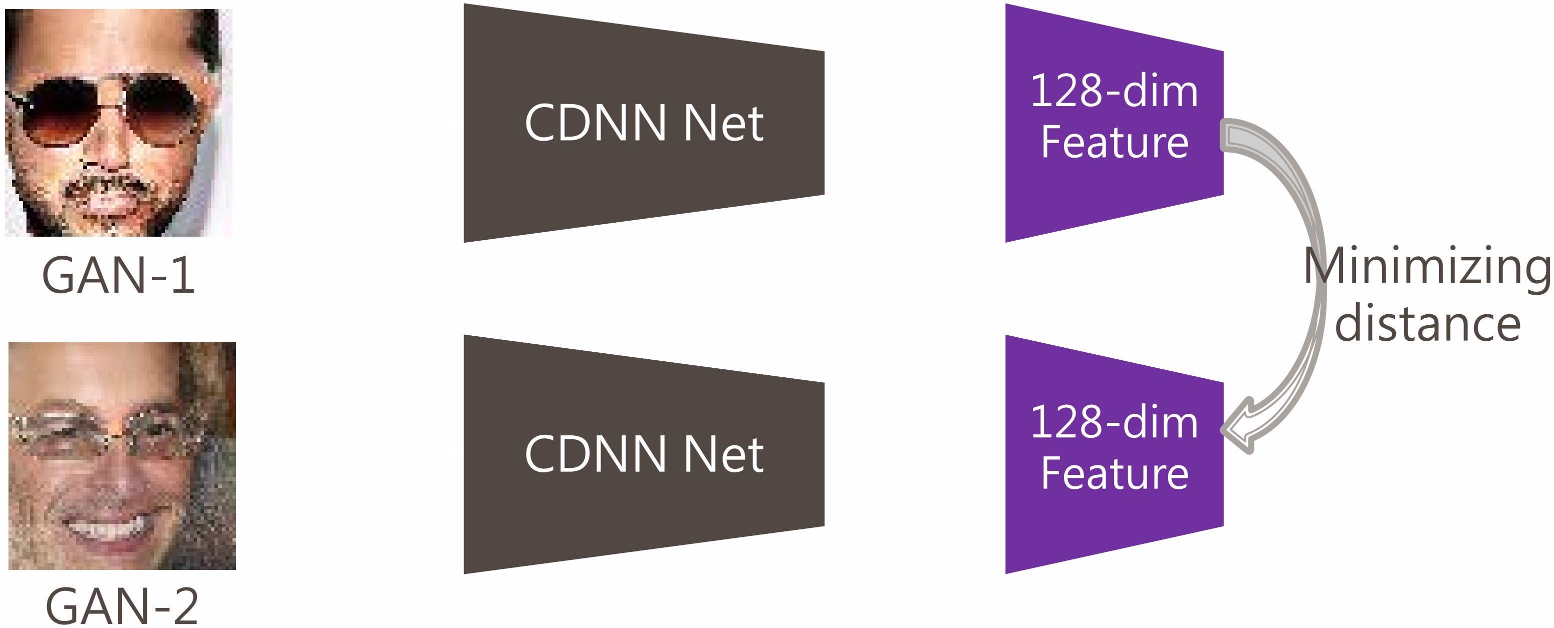
$$\sum_{i}^{N_r}\left[\|\mathrm{D}_1(\mathbf{x}_a)-\mathrm{D}_1\left(\mathbf{x}_p\right)\|_2^2-\|\mathrm{D}_1(\mathbf{x}_a)-\mathrm{D}_1(\mathbf{x}_n)\|_2^2+a\right]_+$$

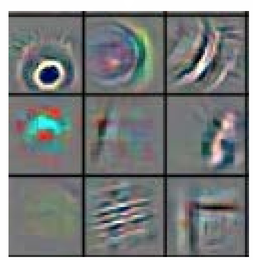


Learning Tricks

- Hard mining is the most important
  - Similar to object detection nets
- Hard positive
  - Same person but different poses in two images
- Hard Negative
  - Different person but looks similar to each other in two images
    - A fake image looks very real
    - A real one looks something wrong
      - May cause by noise or illuminance variantions.

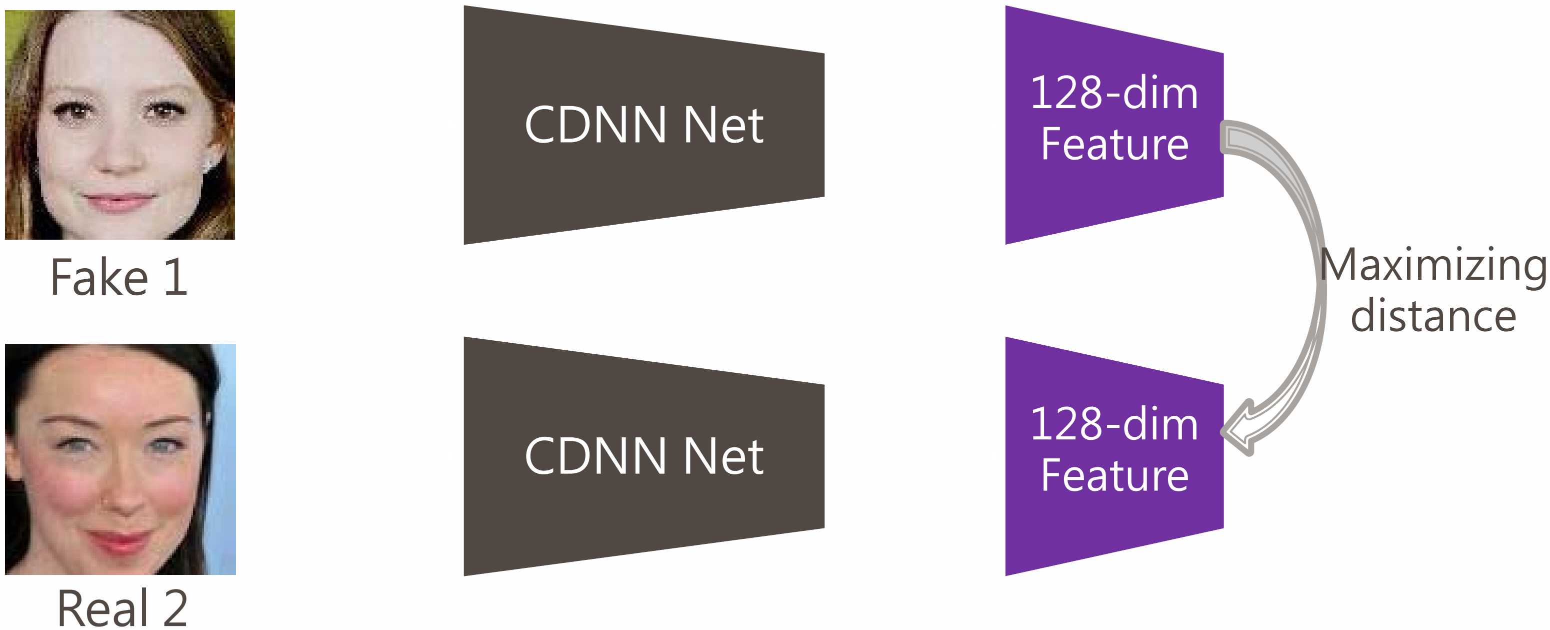
Common Fake Feature Learning

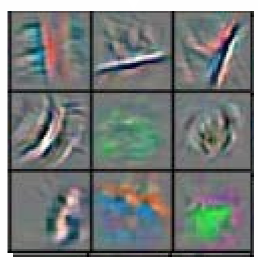




Learning to capture the features of fake images

Common Fake Feature Learning





Learning to capture the features of real images

Classification Network Learning

- Concatenating “traditional classifiers”
  - SVM, Random forest, or Bayer classifier
  - However, we don’t know what features is useful for fake image detection
- Use End-to-end and trainable classifier
  - Learning in supervised way
  - Based on the pre-trained network (CDNN) learned by the proposed pairwise learning

Classification Network Learning

- The loss function of the classifier can be defined as a cross-entropy loss:

$$L_C(\mathbf{x}_i, \mathrm{y}_i) = -\sum_{i}^{N_T} \left(\mathrm{D}_2\left(\mathrm{D}_1(\mathbf{x}_i)\right) \log\ \mathrm{y}_i\right).$$

  - where N_T is the number of the training set and y_i is the label indicating 0 (fake) or 1 (real)

Network Architecture (

Layers	CDNN	Classifier
1	Conv.layer, kernel=7*7, stride=4, channel=96	Conv. layer, kernel=3*3, channel = 2
2	Residual block *2, channel=96	Global average pooling
3	Residual block *4, channel=128	Fully connected layer, neurons=2 Softmax
4	Residual block *3, channel=256	
5	Fully connected layer, neurons=128 Softmax layer	

Experimental Results

- Experimental settings
  - We collect 5 state-of-the-art GANs to generate fake images pool
    - 1) DCGAN (Deep convolutional GAN) [2]
    - 2) WGAP (Wasserstein GAN) [3]
    - 3) WGAN-GP (WGAN with Gradient Penalty) [4]
    - 4) LSGAN (Least Squares GAN) [5]
    - 5) PGGAN [1]
  - Criterion
    - Good quality, different methodologies
  - Each GAN generates 200,000 fake images with sized of 64x64

1. Karras, Tero, et al. "Progressive growing of GANS for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
2. Radford, et al.. "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434,* 2015.
3. M. Arjovsky, et al., "Wasserstein gan," *arXiv preprint arXiv:1701.07875* (2017).
4. Gulrajani, Ishaan, et al. "Improved training of wasserstein gans," *Advances in Neural Information Processing Systems*. 2017.
5. X. Mao, et al. "Least squares generative adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

Experimental Results

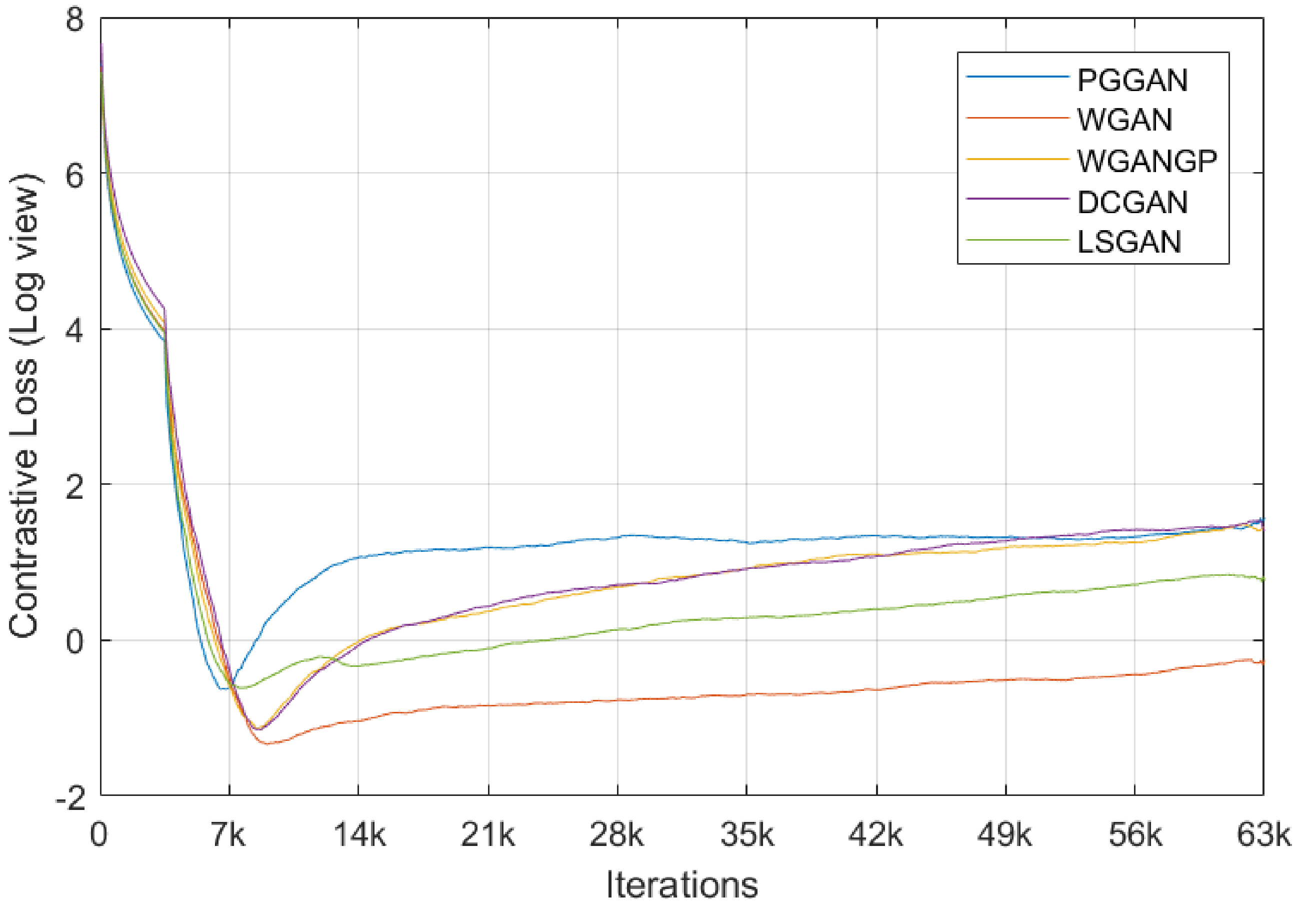
- Experimental settings
  - We randomly pick up 202,599 fake images from the fake images pool
  - Total number of training images: 400,198
  - Total number of test images: 5,000
  - Parameter m in contrastive is 0.5
  - JDF learning in the first two epochs
  - Discriminator learning in the following epochs
- We exclude the fake images generated from one of the collected GANs to verify the proposed method is generalized

Objective Quality Comparison

The performance comparison between the proposed method and other methods

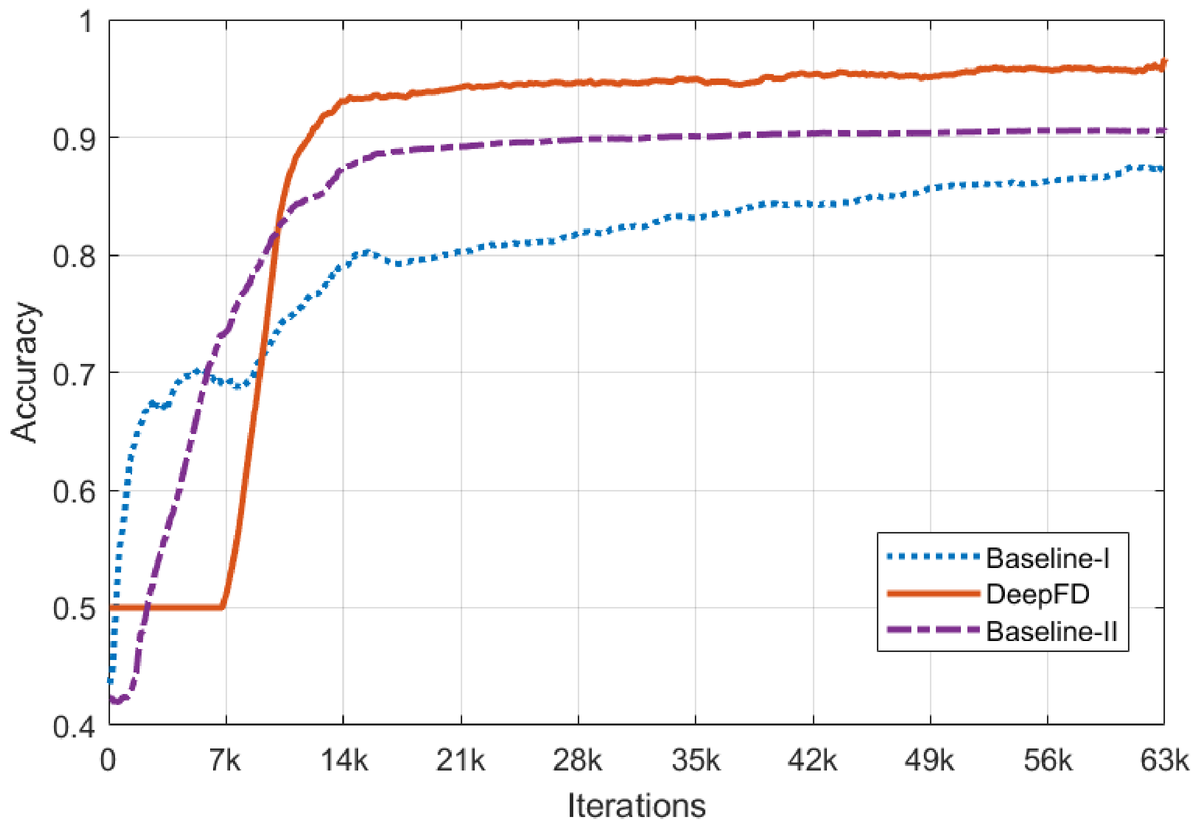
Method/Test target	LSGAN		DCGAN		WGAN		WGAN-GP		PGGAN	
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
Method in [5]	0.205	0.580	0.253	0.774	0.235	0.673	0.242	0.604	0.222	0.862
Method in [7]	0.819	0.528	0.848	0.790	0.817	0.822	0.816	0.679	0.798	0.788
Method in [8]	0.833	0.725	0.812	0.833	0.840	0.809	0.826	0.733	0.824	0.838
Method in [15]	0.947	0.922	0.871	0.844	0.838	0.847	0.818	0.835	0.926	0.918
Baseline-I	0.921	0.915	0.887	0.831	0.860	0.855	0.822	0.837	0.919	0.898
Baseline-II	0.939	0.929	0.878	0.851	0.840	0.863	0.845	0.844	0.922	0.928
Baseline-III	0.845	0.785	0.796	0.816	0.833	0.799	0.819	0.805	0.835	0.854
The proposed	**0.981**	**0.956**	**0.986**	**0.986**	**0.895**	**0.881**	**0.876**	**0.881**	**0.951**	**0.936**

Convergence Analysis of CFF

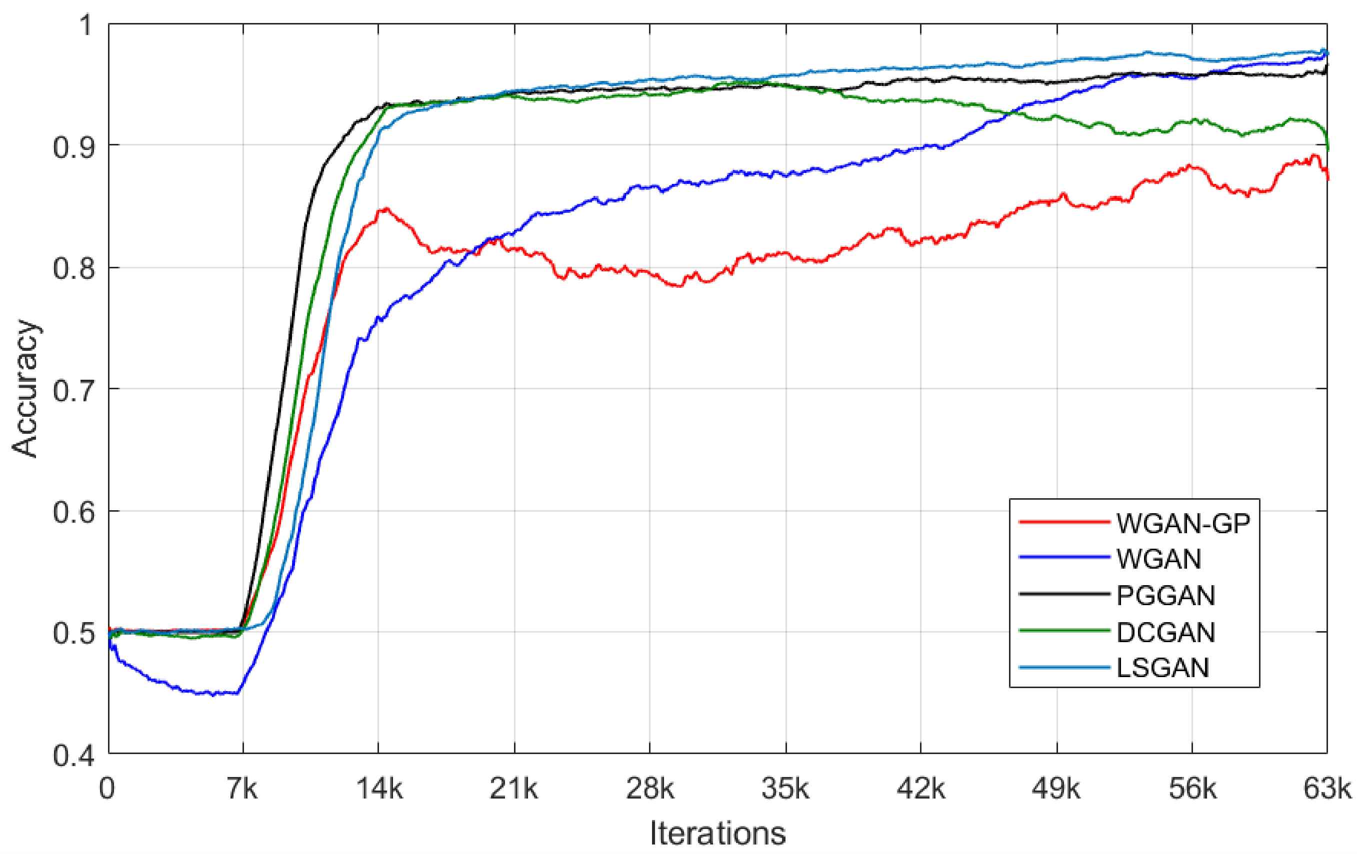


Performance Comparison

- Supervised learning (Baseline-II) vs. pairwise learning



Precision Curves for GANs Used in Our Experiments



Visualized Feature Maps of Fake Image

- Fully convolutional network can be used to visualize the unrealistic details

- (a)-(j): Fake images. (k)-(t) Real ones
- Draw in red indicates fake features.

Conclusion

- The proposed a novel deep forgery discriminator (DeepFD) can successfully detect the fake images
- Contributions
  - The first work to generalize the problems of detecting the fake images
  - The proposed CDNN can capture the common feature for fake images generated by different GANs
  - Visualization of the proposed DeepFD can be used to further improve the detector algorithm

Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

RAM:
RISK ASSESSMENT MODULE FOR AUTONOMOUS DRIVE

許志仲 (Chih-Chung Hsu)

Assistant Professor
Department of Management Information Systems,
National Pingtung University of Science and Technology

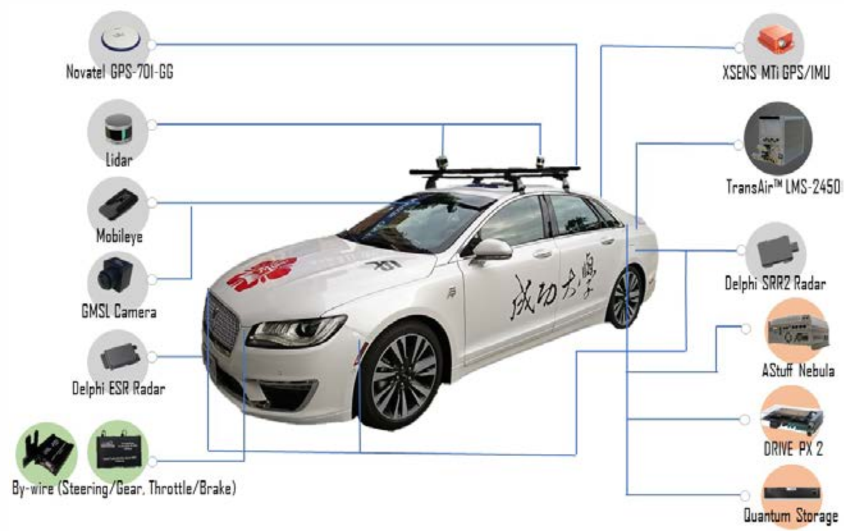
Authors

Chih-Chung Hsu
Assistant Professor
NPUST

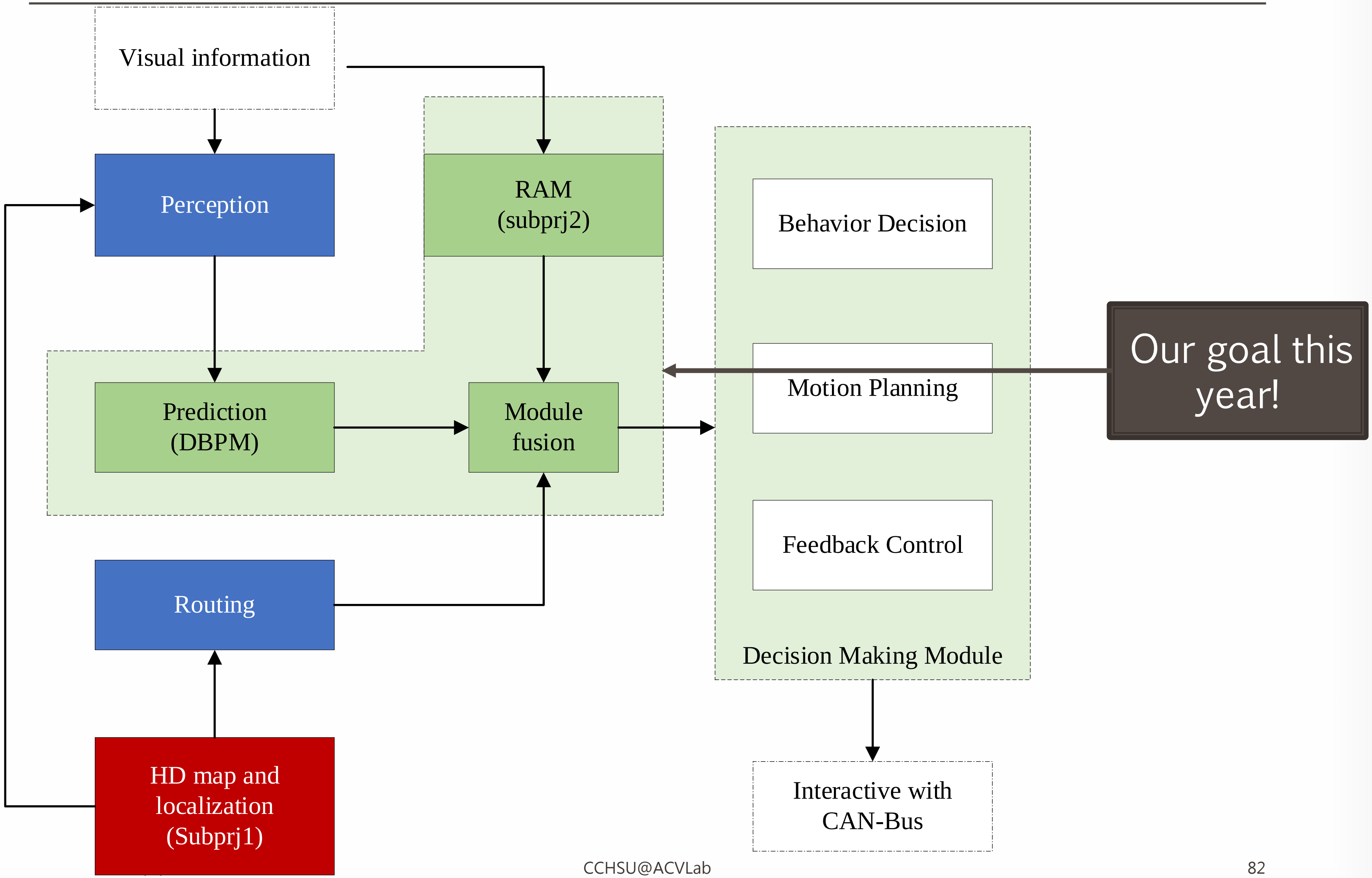
Hao-Ting Yang
Undergraduate Student
NPUST

Wen-Hai Zheng
Undergraduate Student
NPUST

We are running a vision-based ADAS
to detect car accident!
(incorporating with NCKE EE)



Motivation (to have "Risk Assessment Module")



TVCD

- We carefully annotate the objects, especially in collision cases
  - Discover dangerous behaviors for autonomous driving

High-Quality

Low-Quality

Statistics of our TVCD

Duration of original videos		
	#videos	Avg. duration (sec.)
high-risk	150	22.75
Middle-risk	112	22.25
Low-risk	328	20.43
Total	590	21.37

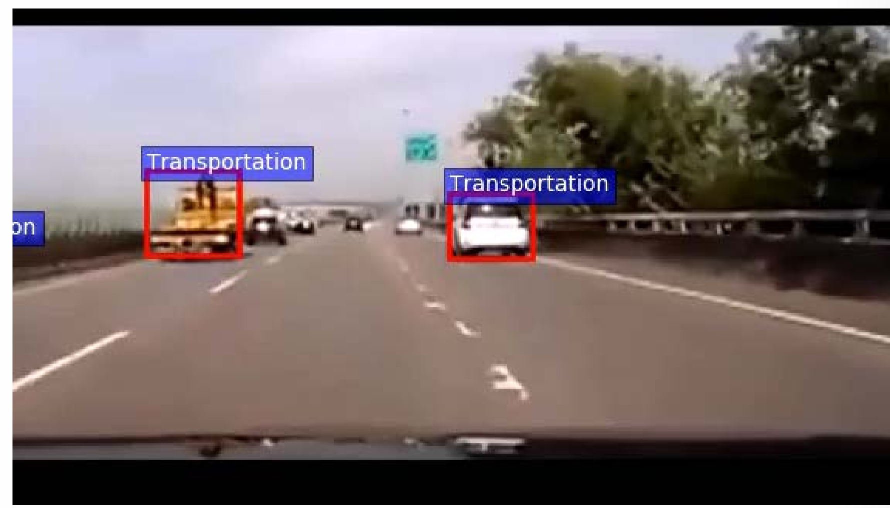
Resolution of original videos		
	MIN	MAX
high-risk	1280*720	1920*1080
Middle-risk	1024*600	1920*1080
Low-risk	800*576	1280*720

Duration of annotated videos		
	#videos	Avg. duration (sec.)
high-risk	75	4.87
Middle-risk	51	4.84
Low-risk	158	4.82
Total	284	4.83

Resolution of annotated videos		
	MIN	MAX
high-risk	1280*720	1280*720
Middle-risk	1280*720	1280*720
Low-risk	1024*600	1024*600

Annotated Samples of TVCD

- The annotated videos will contains
  - Frame-level: Annotations in XML formatted for each frame
  - Video-level: risk-factor, time to accident, and time to out-of-control
  - Normalized resolution/duration involving how car accident occur



SAFE:
SELF-ATTENTION-BASED FEATURE EXTRACTION NETWORK FOR ANTICIPATING DRIVING BEHAVIORS

ICCV 2019 Seoul, Korea

許志仲 (Chih-Chung Hsu)

Assistant Professor
Department of Management Information Systems,
National Pingtung University of Science and Technology

Motivation

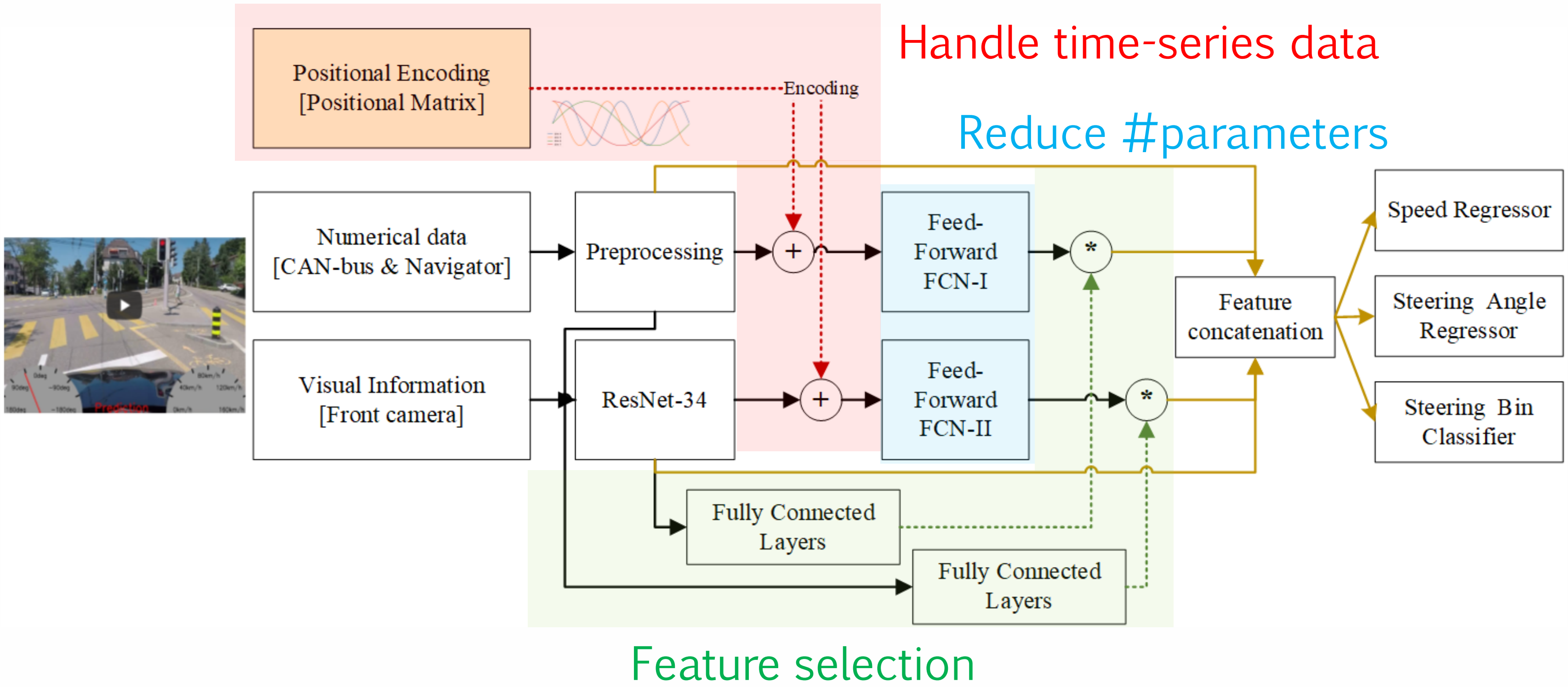
- Autonomous driving system from heterogeneous data
  - Image, maps, CAN-Bus information, etc.
- In real application
  - Fast and accurate prediction is required
    - Inference complexity should be minimized
  - L2D challenge is one of our project’s goals

Predict the future

- Handling time-series data, recurrent network is widely used
  - LSTM / GRU etc.
  - Pros:
    - Capture temporal information well
  - Cons:
    - Hard to parallel processing
- We may not care "training complexity" but inference complexity
  - A feed-forward CNN for driving behavior prediction is proposed
    - Inspired by Transformer in NLP, we have designed 3 key components
      1. *Positional encoding*
      2. *Fully convolutional neural network for extracting feature*
      3. *Self-attention mechanism*

Proposed Method

- Training flowchart of the proposed Self-Attention-based Feature Extraction Network (SAFE)

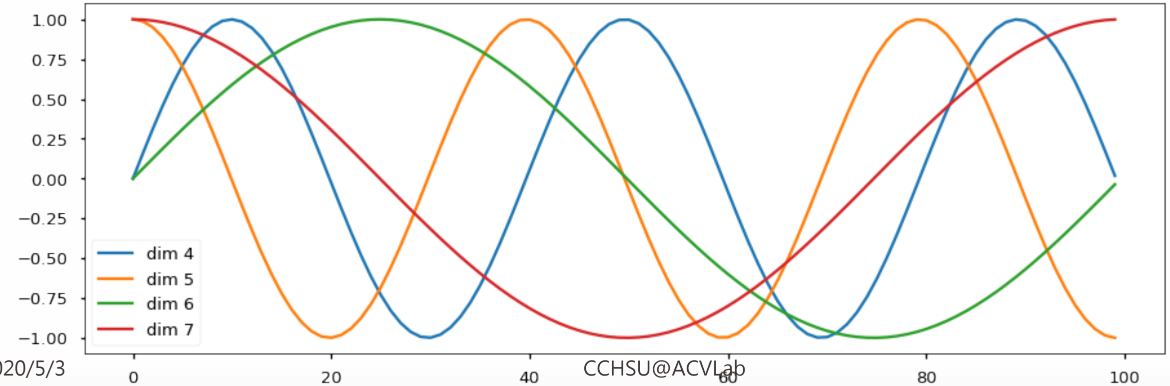


Handling Time-Series in CNN

- We removed the recurrent networks in the SAFE
  - To capture the temporal information
    - Positional encoding is required

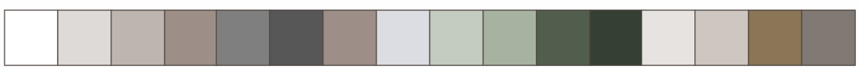
$$PE_{(t,2i)} = \sin\left(\mathrm{t}\ /10000^{2i/d_m}\right)$$
$$PE_{(t,2i+1)} = \cos\left(\mathrm{t}\ /10000^{2i/d_m}\right),$$

    - where the *t* is the feature at time *t*, *dm* is the number of dimensions of given feature, and *i* indicates *i*-th dimension in the given feature vector.



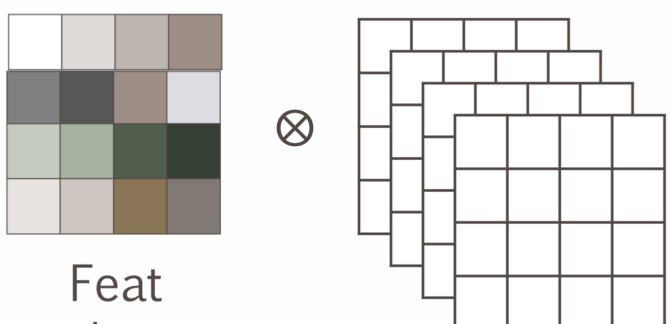
FCN Feature Extraction

- A lot of #parameters in fully connected layer
  - Keeping feature correlation as well as reduce #parameters
    - CNN is used to capture CAN-Bus data



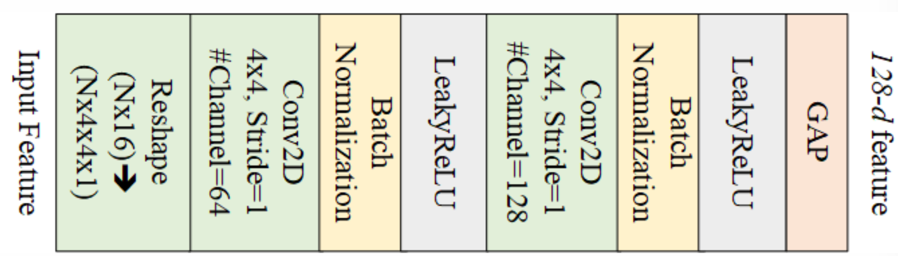
2FCs: 16*128+128*128=18432

2Convs: 4*4*128+4*4*128=4096



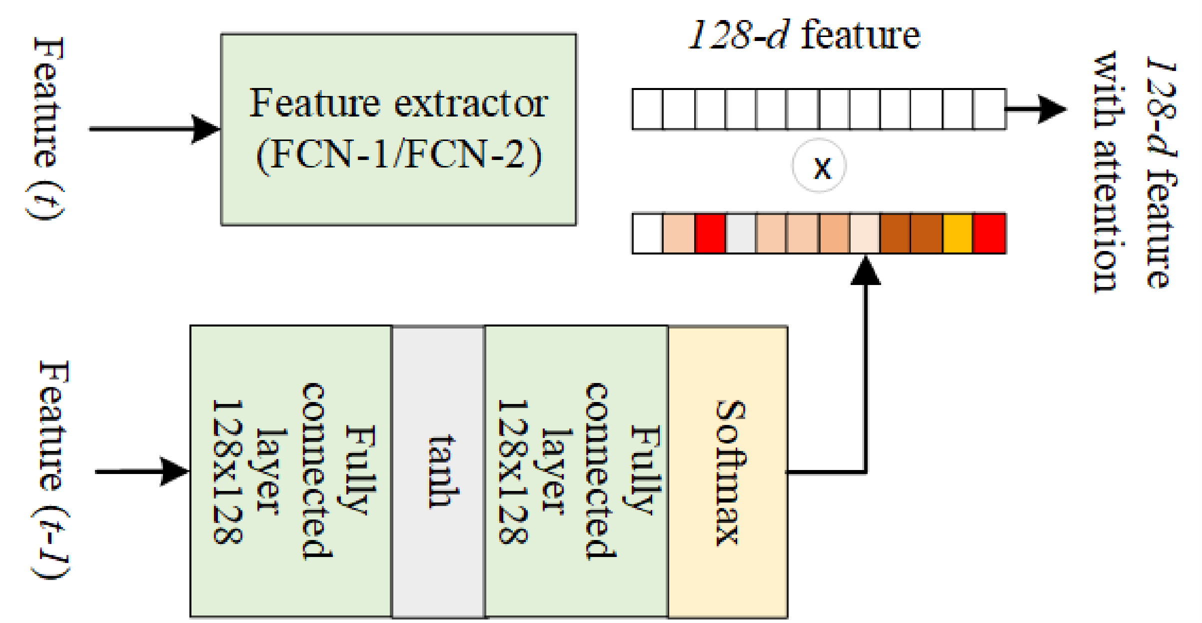
Feat reshape

k-channeled 4x4 conv.



Finding Important Features

- Self-Attention mechanism
  - Capture the information from time *t*
    - FC + Softmax ➔ attention
  - Feature reweighting at time *t*+1 by the attention calculated from feature at time *t*.



Experimental Result

- We use train/validation sets provided by the challenge
  - Information used in our SAFE model
    - Front camera video only (resize to 224x224)
    - CAN-Bus information (16 features)
  - Since our hardware is limited, we only adopt parts information from original data.
    - GTX1080Ti *1 + i7-7700 + 16G RAM.
- Training tricks
  - Higher weight for losses for steering angle predictors in first 10 epochs
  - Fine-tuning by equivalent weights

Ablation Study

- We have tested each part of the proposed SAFE model to verify its effectiveness
  - SAFE-I: SAFE model without self-attention mechanism.
  - SAFE-II: SAFE model without positional encoding.
  - SAFE-III: Recurrent network is used to capture temporal information (we use GRU) and SAFE model without positional encoding.
  - SAFE-IV: SAFE model without FCN sub-networks (say, use fully connected layers instead).

Result

Table 1. Performance comparison among our SAFE model with different settings.

Method	MSE_A	MSE_S	CE_A
RF	0.381	0.311	0.841
SAFE-I	0.207	0.151	0.628
SAFE-II	0.221	0.170	0.633
SAFE-III	0.195	0.181	0.621
SAFE-IV	0.199	0.177	0.625
SAFE	**0.175**	**0.153**	**0.589**

- The inference time of our SAFE model
  - 26 fps without code optimization
- SAFE-III (GRU instead)
  - 9 fps.

Conclusion

- We have proposed SAFE model which
  - Can effectively capture temporal information in a feed-forward network
  - Reduces the #parameters while keeping the performance
- We still working on it and two different approach will combine our SAFE model
  - Car accident prediction
  - Dangerous driving behaviors analysis

Research Highlights

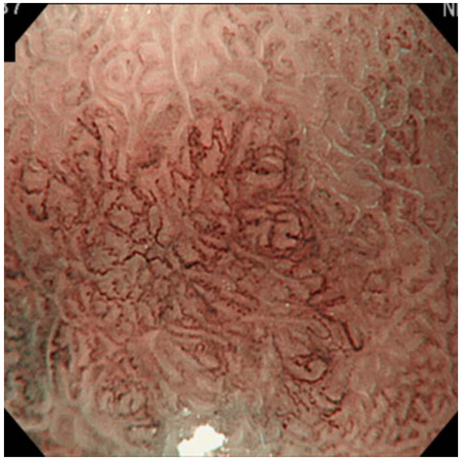
- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

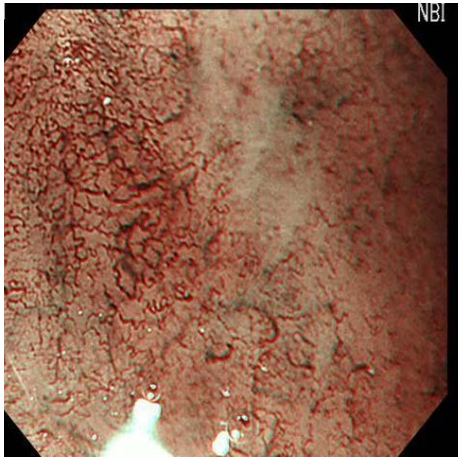
SSSNET: SMALL-SCALE-AWARE SIAMESE NETWORK FOR GASTRIC CANCER DETECTION

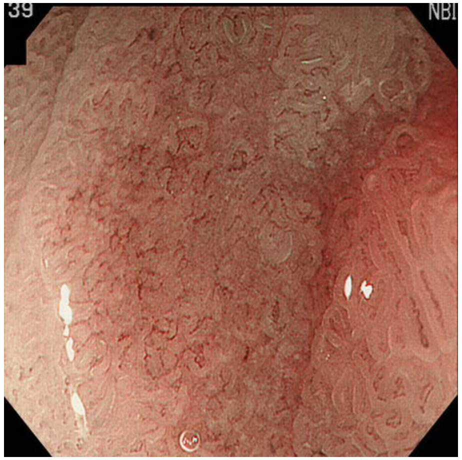
IEEE AVSS’ 19, Oral
Contribute to MOST-AI Project (NTHU)

Introduction

- Detection of early gastric cancer cells by M-NBI technology



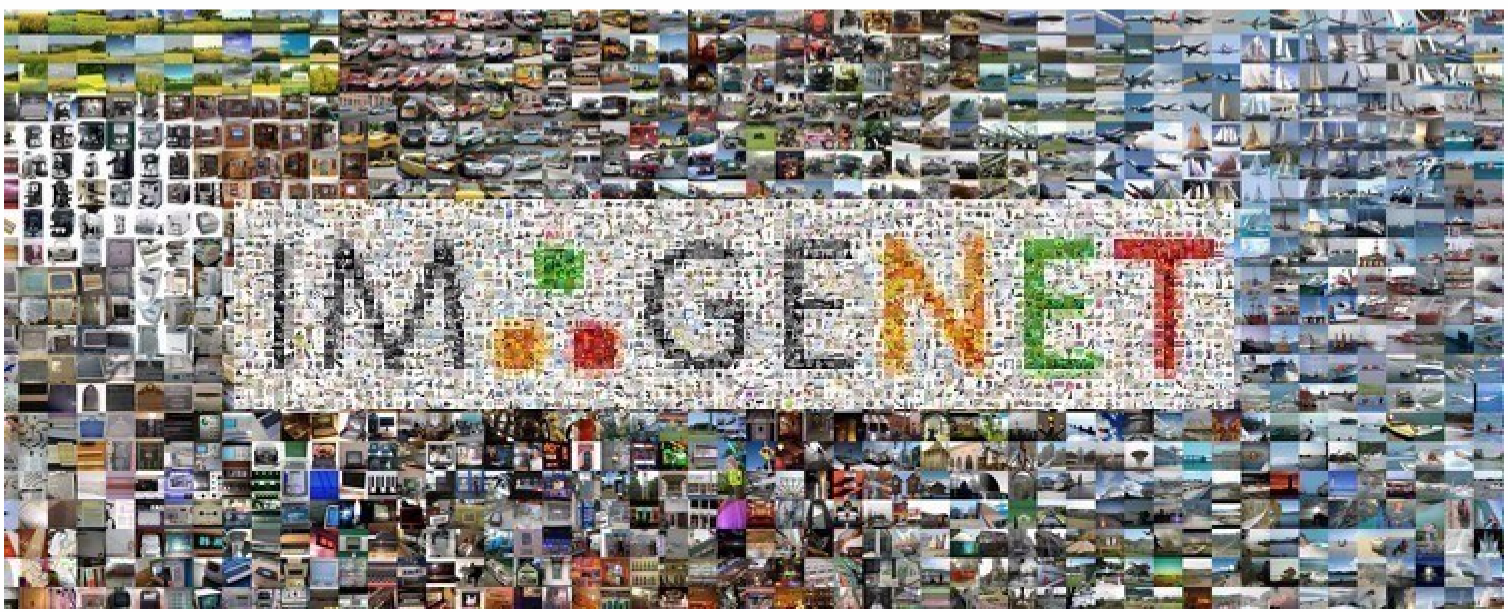




Cancerous	Cancerous	Normal

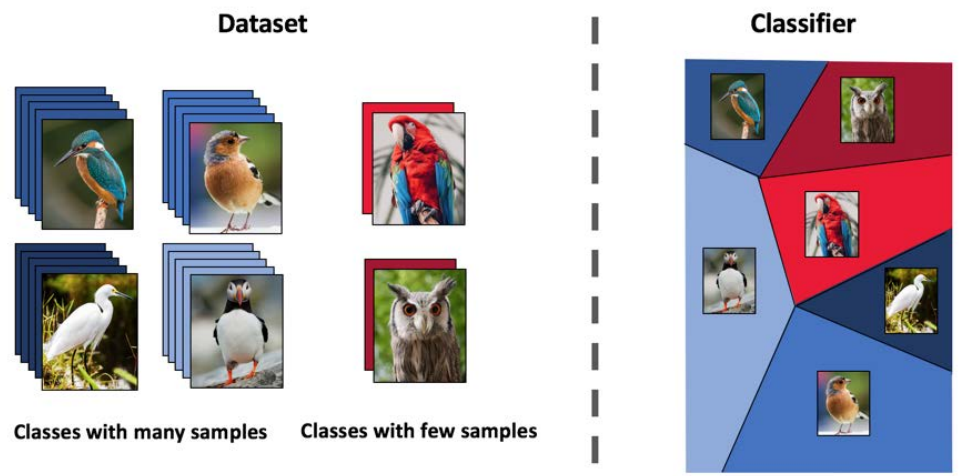
Motivation

- #Medical images is limited
  - Transfer learning is hard to used in this case
- Small scale training sets ➔ overfitting
  - Neural network architecture should be simplified



Related Work

- Few-Shot Learning
  - Model-based [1]
    - Transfer learning, domain adaptation
  - Metric-based [2]
    - Siamese network based
  - Optimization approach [3]



1.Binford, Thomas O. "Survey of model-based image analysis systems." *The International Journal of Robotics Research* 1.1 (1982): 18-64.
2.Ferzli, Rony, and Lina J. Karam. "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)." *IEEE transactions on image processing* 18.4 (2009): 717-728.
3.Afonso, Manya V., José M. Bioucas-Dias, and Mário AT Figueiredo. "Fast image recovery using variable splitting and constrained optimization." *IEEE transactions on image processing* 19.9 (2010): 2345-2356.

Our Method

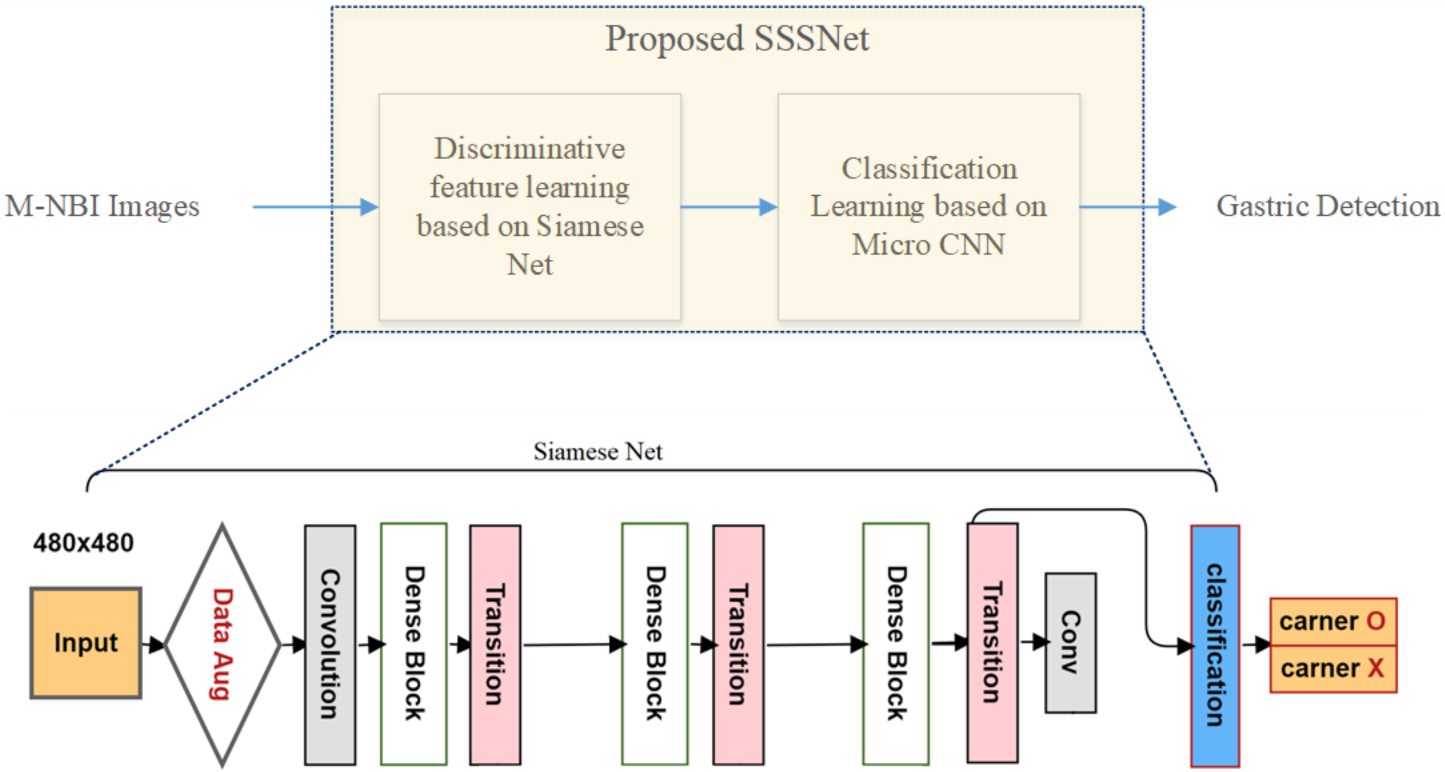


Figure 1. The proposed method including SSSNet and learning policy.

Method based on Contrastive Loss

- Based on pairwise learning to learn the discriminative feature first

$$E_W(x_1, x_2) = ||f(x_1) - f(x_2)||_2^2$$

$$L\big(W, (P, x_1, x_2)\big) = 0.5 \times \left(y_{ij} E_w^2\right) + (1 - y_{ij}) \times max(0, (m - E_w\ )_2^2)$$

Method (Fine-tuning Phase)

- Learning a classifier by cross-entropy

$$L_c(x_1, p_1) = -\sum_{i}^{N_T} (f_{cls}\big(f_{sia}(x_1)\big) log p_i)$$

- The total loss function will be

$$L(x_1, x_2, p_1, y_1) = \alpha L_c(x_1, p_1) + (1-\alpha) L(W, (P, x_1, x_2))$$

  - where α is a balance factor
    - $\alpha = 0$ for the first 10 epochs
    - $\alpha = 0.4$ for the rest

Experiment Setting

- Data classification

Data	images
Typical case	130
Difficult case	343

- Data splitting

Training	400
Validation	13
Test	60

- Training settings

lr	1e-3
Epochs	60
Optimizer	Adam

Experimental Result

Table 1. Comparison of detection rate evaluated for the proposed method and other baselines.

Method	Precision	Recall	Specificity	Accuracy	F-measure
DenseNet-12	0.417	0.385	0.500	0.444	0.400
ResNeXt	0.500	0.462	0.571	0.519	0.480
EffcientNet	0.429	0.462	0.429	0.444	0.444
MobileNet v3	0.467	0.538	0.429	0.481	0.500
Baseline-1	0.815	0.838	0.779	0.810	0.826
Baseline-2	0.462	0.462	0.500	0.481	0.462
SSSnet(proposed)	0.934	0.900	0.937	0.918	0.917

Conclusion

- Based on :
  - Siamese network
  - DenseNet
- SSSNet architecture can be used to learn the discriminative feature from a small-scale training set effectively
- Can improve the performance of gastric cancer detection in M-NBI images.

Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

STRONGER BASELINE FOR VEHICLE RE-IDENTIFICATION

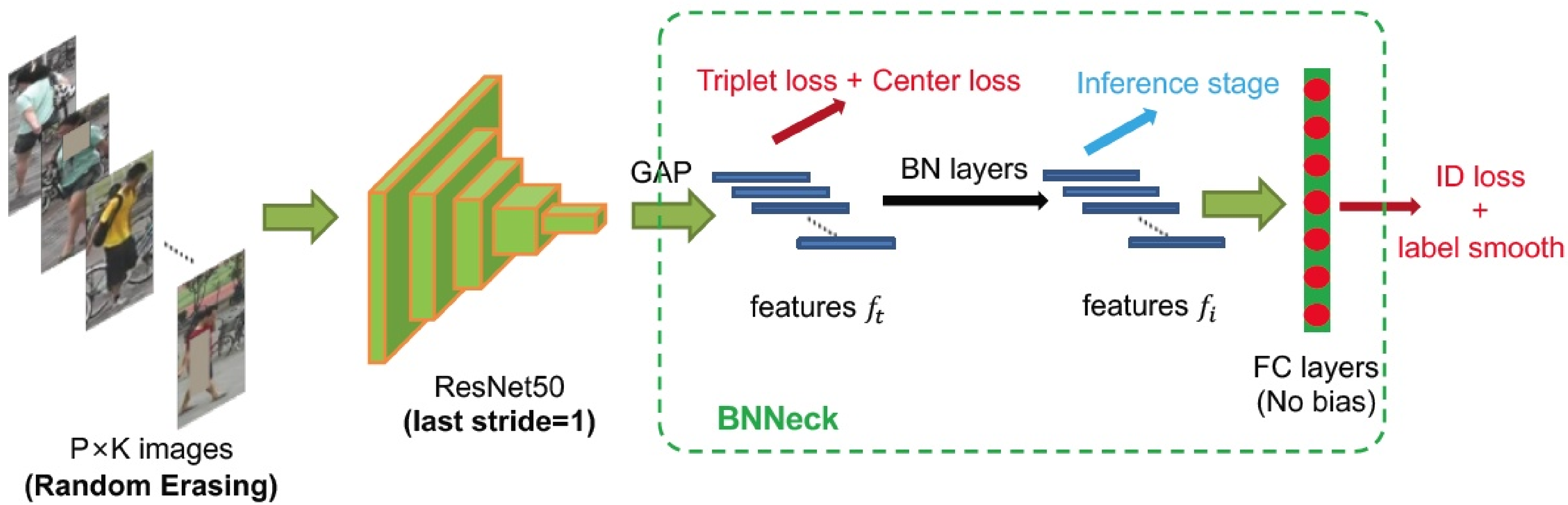
VCIP19'
3rd place, Grand Challenge on Vehicle Re-identification in the wild
Contribute to my MOST project

Vehicle/Person Re-Identification (ReID) Tasks

- Given a query image
  - Find the image(s) with the same identity with the query image
  - Discriminative feature is necessary

SOTA in ReID

- It is common way to learn the discriminative feature based on contrastive and triplet loss functions
- Current SOTA: Strong baseline
  - Bigger feat map + center loss



Luo, Hao, et al. "Bag of tricks and a strong baseline for deep person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

Strong Baseline for ReID

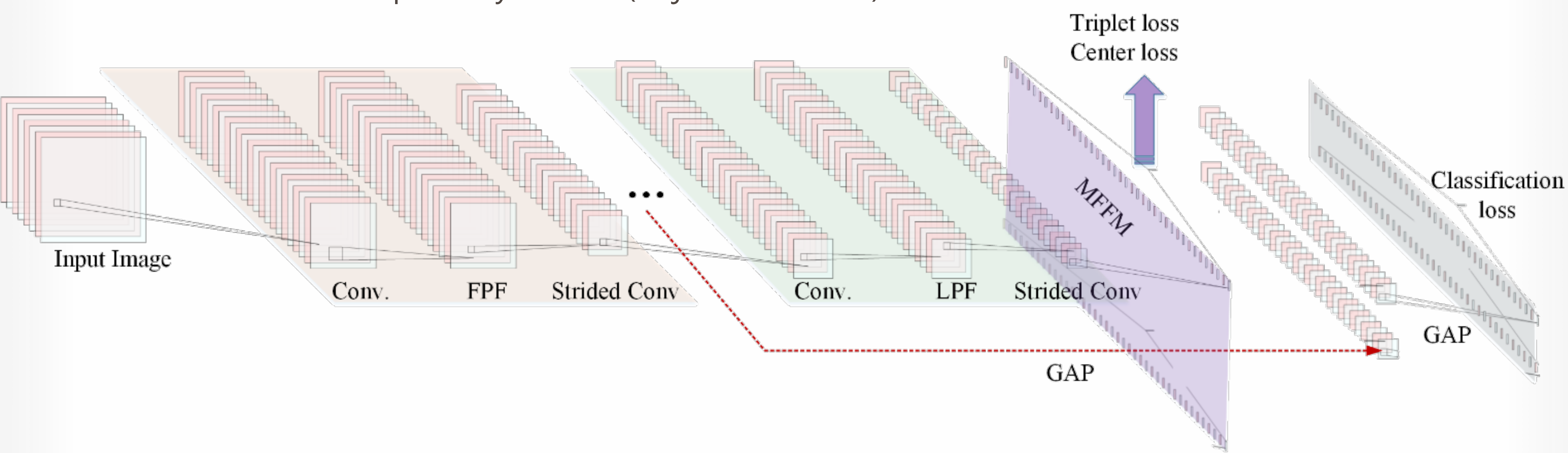
- SOTA in person/vehicle ReID tasks
  - The dataset is contracted in a controllable environment

- Shortcomings:
  - ResNet-50 backbone: not powerful now
  - Not verified in a real-world dataset
    - Vehicle ReID dataset in the wild [1]
  - No cross-layer feature maps are used

[1] Lou, Yihang, et al. "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

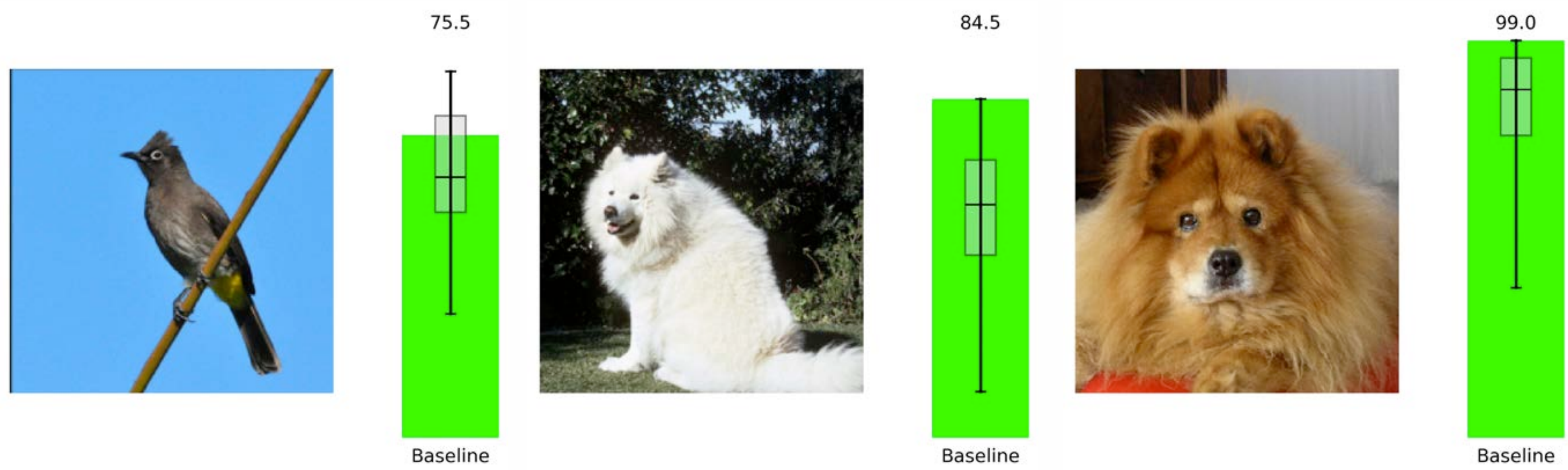
Proposed Stronger Baseline for ReID

- A good baseline leads to good performance in ReID
  - We have integrated
    - Anit-aliasing CNN
      - Proposed by Adobe Research (ICML19)
    - Multi-layer Feature Fusion Module (MFFM)
      - Inspired by M2Det (object detection)



Deep Networks are not Shift-Invariant

Accuracy vary when shifting pixels



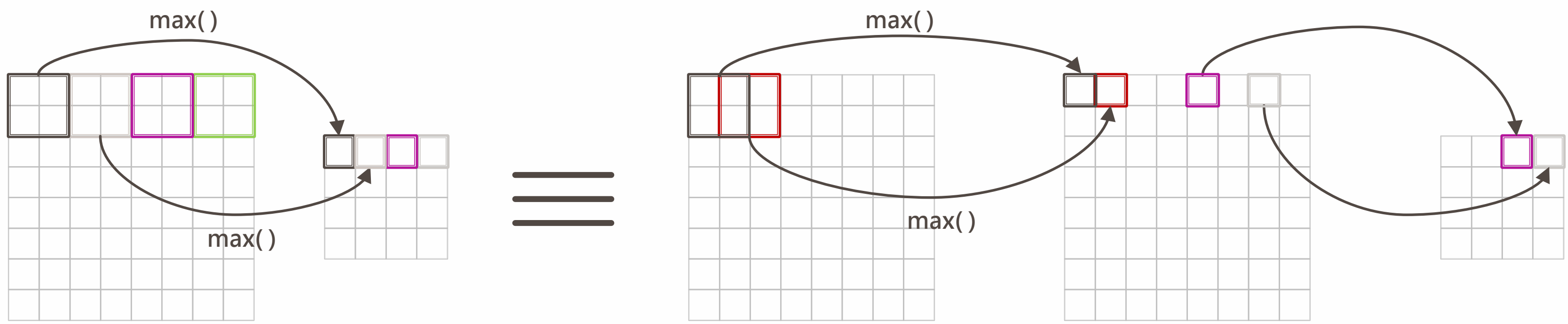
Azulay and Weiss. **Why do deep convolutional networks generalize so poorly to small image transformations?** In ArXiv, 2018.
Engstrom, Tsipras, Schmidt, Madry. **A rotation and a translation suffice: Fooling cnns with simple transformations.** In ArXiv, 2017.

But why?

- Convolutions are shift-equivariant
- Pooling builds up shift-invariance
  - Max pooling
  - Strided convolution

- Anti-aliasing?
  - Blurring before downsampling
    - Basic concept in [1]

[1] Adrian Davies and Phil Fennessy (2001). *Digital imaging for photographers* (Fourth ed.). Focal Press. ISBN 0-240-51590-0.



Strided Pooling

Shift-equivariance lost; heavy aliasing

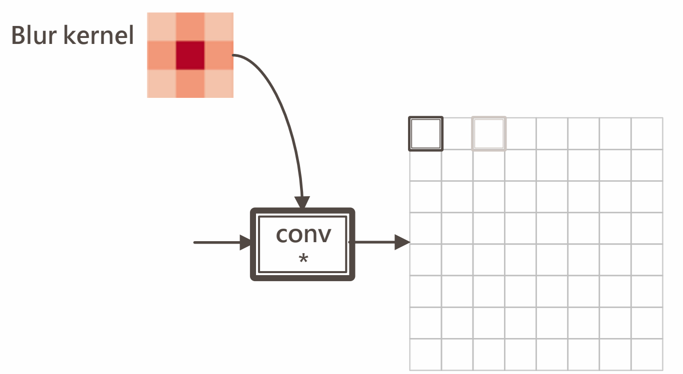
Max (densely) + Subsampling

Preserves shift-equivariance

Shift-eq. lost; heavy aliasing

MaxPool

Equivalent Interpretation



\+ Blur

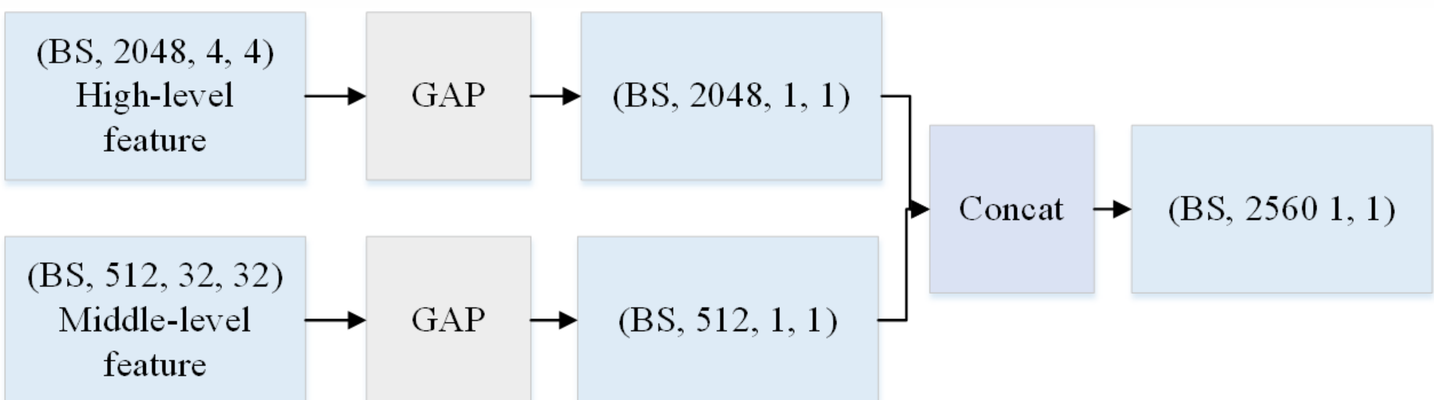
Preserves shift-eq.

Shift eq. lost, but with reduced aliasing

Anti-aliased pooling (MaxBlurPool)

Multi-layer Feature Fusion Module (MFFM)

- We adopt middle- and high-level features as our base feature for ReID
  - To better preserving the spatial information
    - We adopt global averaging pooling instead of fully connected layer



Experimental Results

- Dataset: ReID-Wild
  - Dataset
    - 416,314 vehicle images with 40,671 identities
  - Training set:
    - 380,000 images with 40,671 identities
  - Validation set:
    - 36,314 images with 40,671 identities
  - Testing:
    - Small: 3,000 identities with 38,862 images
    - Middle: 5,000 identities with 64,390 images
    - Large: 10,000 identities with 128,518 images

Experimental Results

Methods	Small	Middle	Large
GoogLeNet [12]	24.27	24.15	21.53
Triplet [13]	15.69	13.34	9.93
Softmax [14]	26.41	22.66	17.62
CCL [15]	22.50	19.28	14.81
HDC [16]	29.14	24.76	18.30
GSTE [17]	31.42	26.18	19.50
UGAN [18]	29.86	24.71	18.23
EN [7]	28.77	24.63	19.48
FDA w/ At [7]	32.40	27.10	21.13
FDA [7]	35.11	29.80	22.78
BTSB [4]	39.61	33.24	28.98
Proposed	**51.38**	**43.61**	**37.91**

mAP (Mean Averaging Precision) comparison

Top-k Accuracy Comparison

Method	Small		Middle		Large	
	$R1$	$R5$	$R1$	$R5$	$R1$	$R5$
GoogLeNet [12]	57.16	75.13	53.16	71.1	44.61	63.55
Triplet [13]	44.67	63.33	40.34	58.98	33.46	51.36
Softmax [14]	53.4	75.03	46.16	69.88	37.94	59.89
CCL [15]	56.96	75.0	51.92	70.98	44.6	60.95
HDC [16]	57.1	78.93	49.64	72.28	43.97	64.89
GSTE [17]	60.46	80.13	52.12	74.92	45.36	66.5
UGAN [18]	58.06	79.6	51.58	74.42	43.63	65.52
EN [7]	57.13	77.33	52.86	73.18	43.02	66.3
FDA w/ At [7]	61.93	80.48	55.62	75.64	46.48	68.36
FDA [7]	64.03	82.8	57.82	78.34	49.43	70.48
BTSB [4]	71.73	85.53	66.5	81.65	60.59	76.77
Proposed	**82.73**	**92.53**	**78.26**	**91.84**	**71.18**	**87.41**

Ablation Study

- Baseline-I: Proposed method without anti-aliasing
- Baseline-II: Proposed method without MFFM

Top-k Accuracy Comparison

Method	Small		Middle		Large	
	$R1$	$R5$	$R1$	$R5$	$R1$	$R5$
Baselin-I	75.15	84.61	68.1	83.42	63.71	79.91
Baselin-II	76.33	86.71	70.71	85.75	65.33	82.64
BTSB [4]	71.73	85.53	66.5	81.65	60.59	76.77
Proposed	**82.73**	**92.53**	**78.26**	**91.84**	**71.18**	**87.41**

Top-k Accuracy Comparison

Methods	Small	Middle	Large
Baselin-I	41.22	34.63	29.41
Baselin-II	42.37	38.56	32.64
BTSB [4]	39.61	33.24	28.98
Proposed	**51.38**	**43.61**	**37.91**

Conclusion

- Main contribution
  - Stronger baseline
    - Multi-layer feature fusion is effective
    - Shift-invariant (anti-aliasing) CNN can capture better visual features
  - We have won the 3rd place in VCIP grand challenge
    - Only 3 days to train

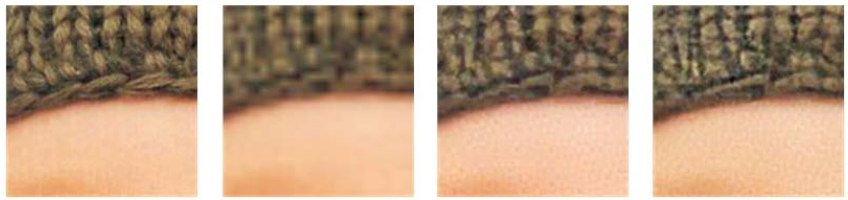
Research Highlights

- Overview of Deep Learning
  - Supervised – Unsupervised – Semi-supervised Learning
- Pairwise Learning based Applications
  - Identity-preserving face hallucination [18-19]
  - Fake face image detection [18-]
  - Risk assessment module for autonomous car [19-]
  - Vehicle Re-identification in the wild [19-]
  - Gastric cancer detection for small-scale M-NBI dataset [19-]
- Other computer vision applications
- Summary

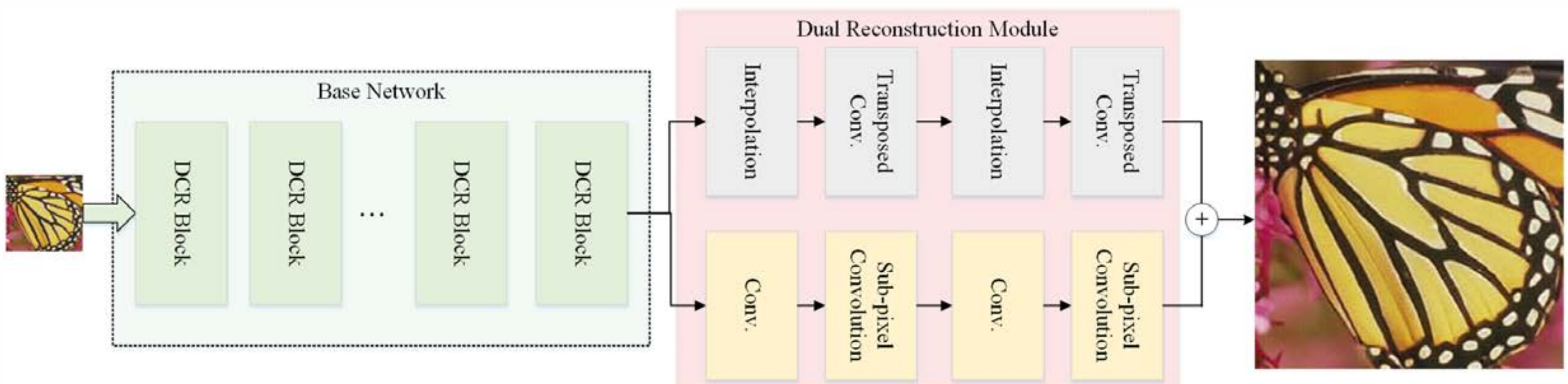
DUAL RECONSTRUCTION WITH DENSELY CONNECTED RESIDUAL NETWORK FOR SINGLE IMAGE SUPER-RESOLUTION

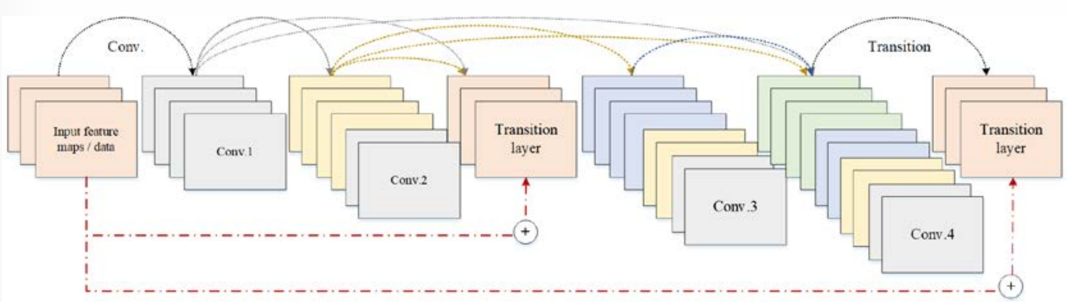
ICCV 2019, Workshop on Advances Image Manipulation
5nd place in Single Image Super-Resolution Challenge (ICCV)

Our Dual Reconstruction Method (9 days)



HR　　Bicubic　　ESRGAN　　OURS





Results (Winner in the Challenge)

SRC: Rank-Correlation
MSE: Mean Squared Error
MAE: Mean Absolute Error

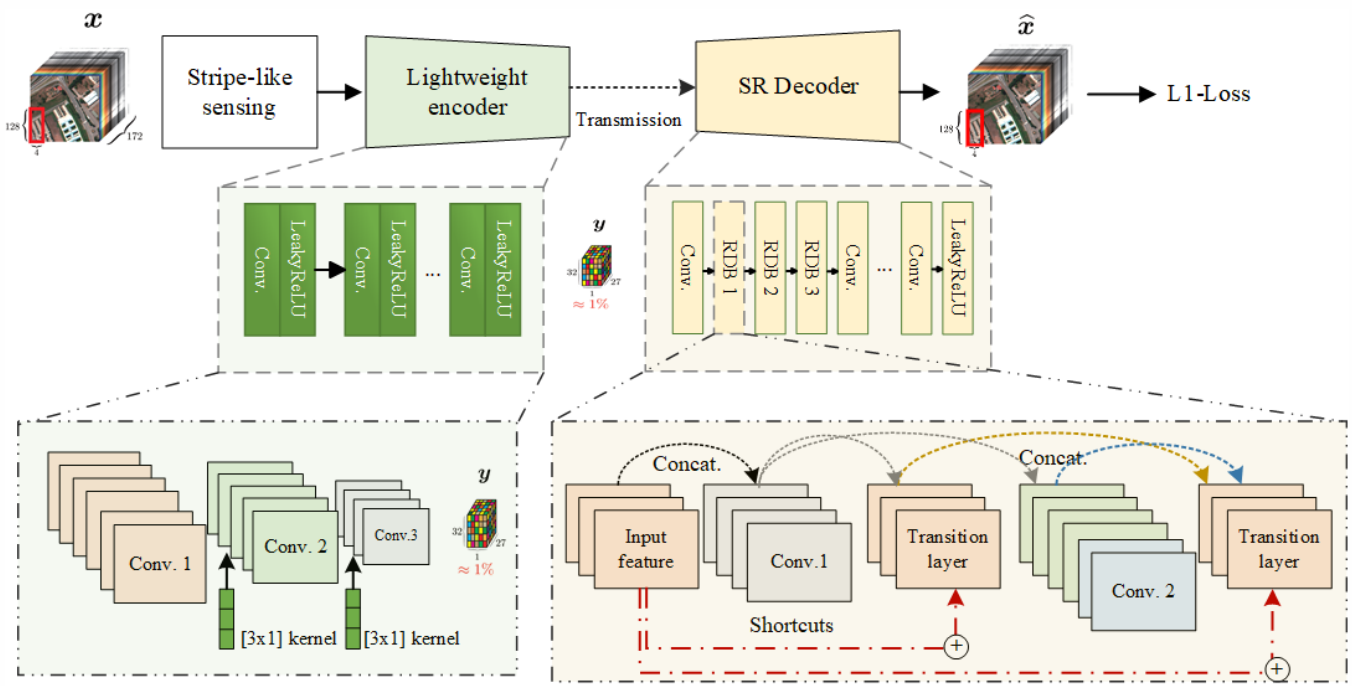
Methods	SRC	MSE	MAE
Baseline-I	0.448	7.595	2.107
Baseline-II	0.450	5.411	1.846
Baseline-III	0.461	5.068	1.785
Baseline-IV	0.470	5.442	1.871
MM [5]	0.528	5.891	1.942
IR [4]	0.537	5.872	1.939
EW [8]	0.548	5.856	1.938
Proposed w/o text-based data	0.376	5.049	1.810
Proposed w/o image data	0.622	3.993	1.588
Proposed w/o numerical data	0.611	3.940	1.552
Proposed	**0.656**	**3.561**	**1.497**

DEEP HYPERSPECTRAL COMPRESSIVE SENSING

Preparing (with Prof. Chia-Hsiang Lin, NCKU EE)

Deep Compressive Sensing

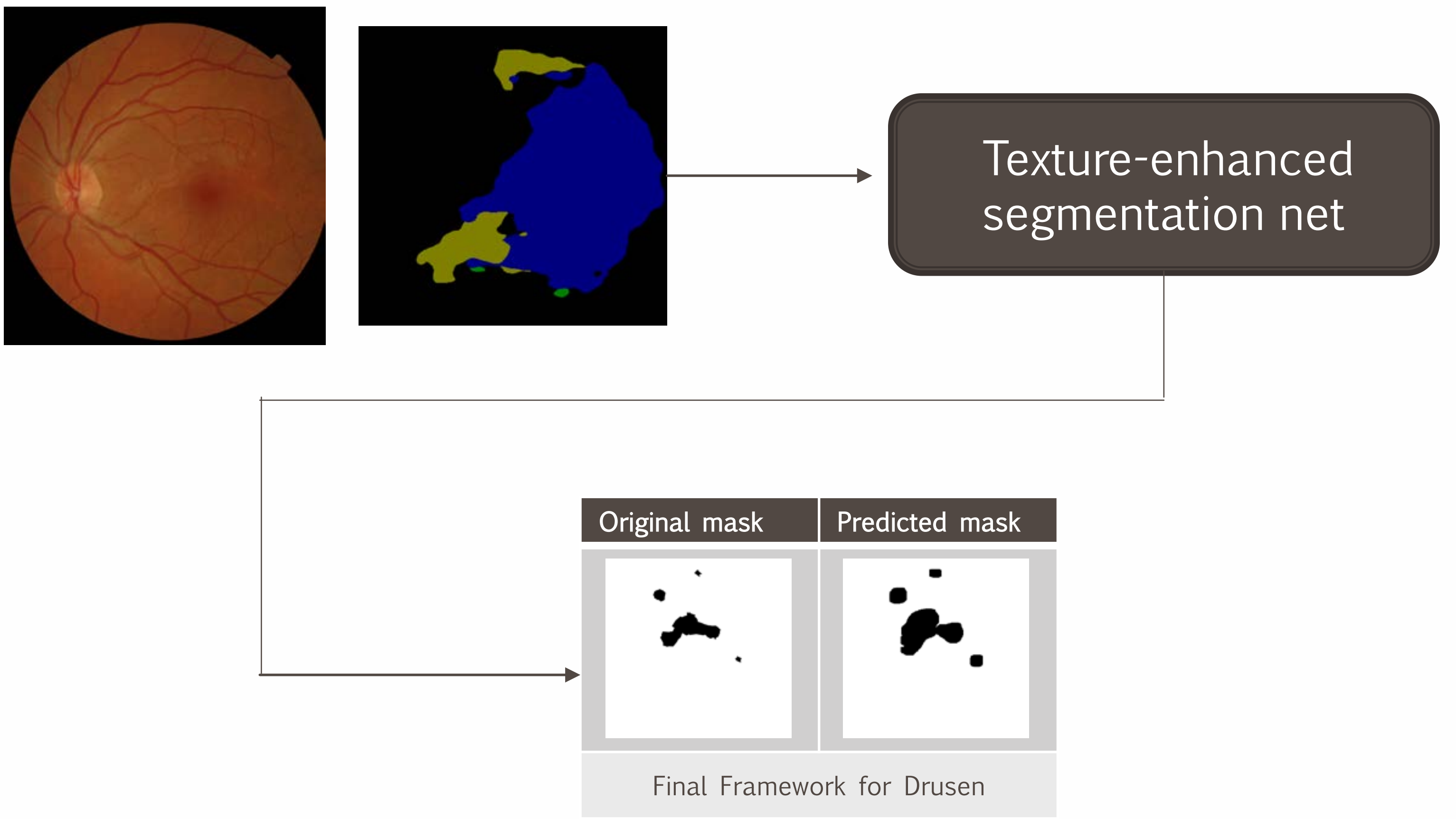
- Very fast sensing, accurately reconstructing, and compressively.
  - For miniaturized satellites



DETECTION AND SEGMENTATION OF LESIONS FROM FUNDUS IMAGES

Preparing
3rd Place, ADAM Challenge, IEEE ISBI Conference (Top-conference on medical image processing)

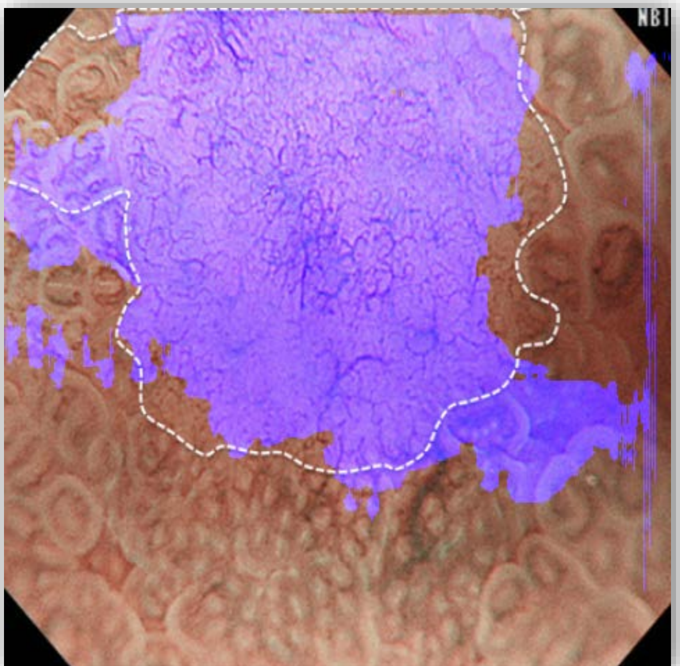
Novel Segmentation Network

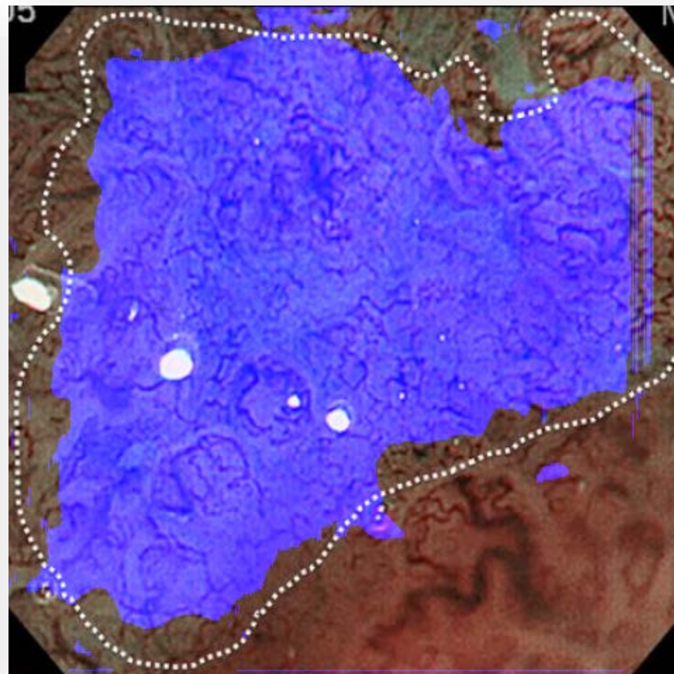


GASTRIC DETECTION FOR M-NBI

AI.SKOPY, 2018
USA Patent

Real-time Cancerous Detection @ 90% Precision

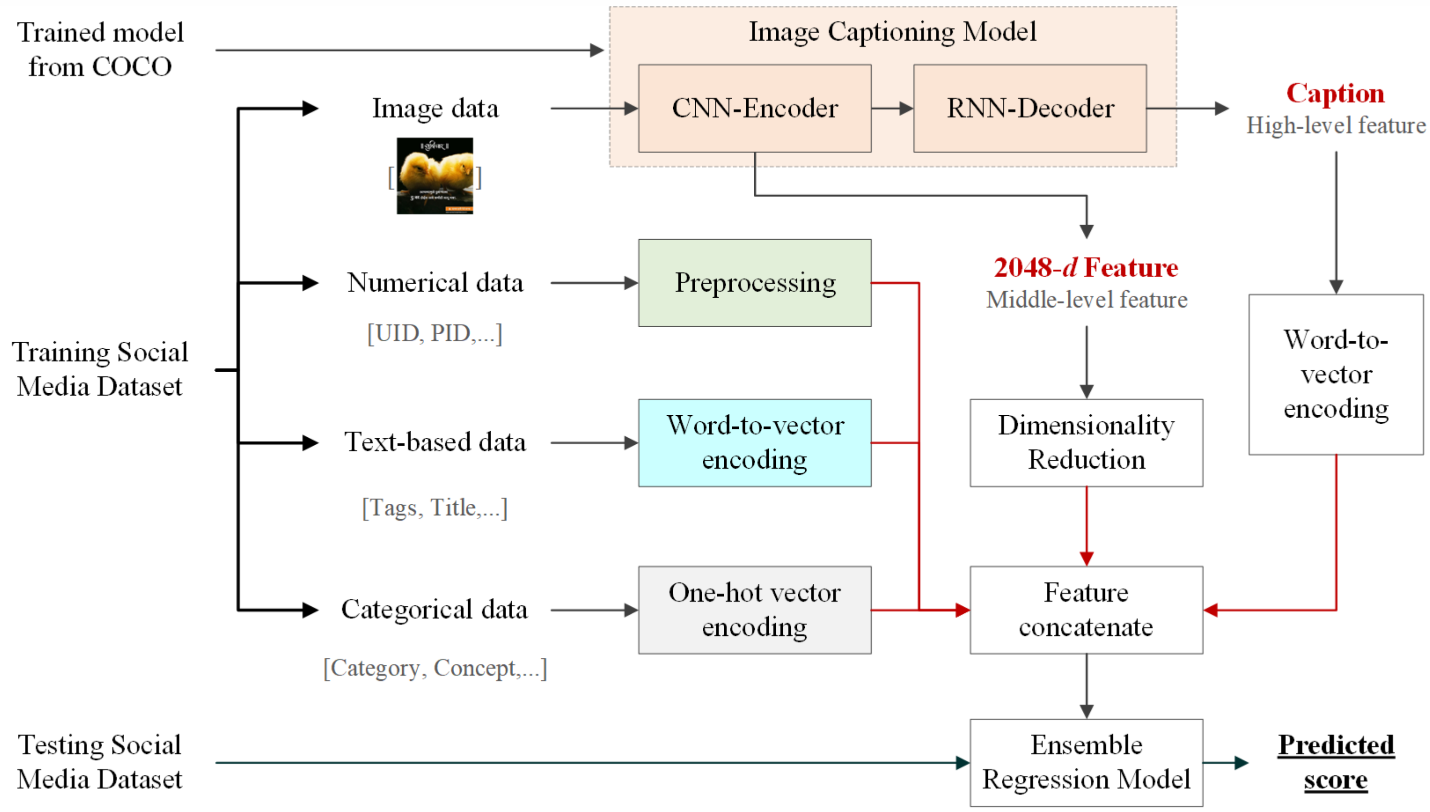




POPULARITY PREDICTION OF SOCIAL MEDIA BASED ON MULTI-MODAL FEATURE MINING

ACMMM 19
Winner in Social Media Prediction Challenge (ACMMM)

Our Multi-modal Feature Mining Method



Conclusion

- Pairwise learning is useful in various tasks
  - More and more attraction about "contrastive coding"
    - Based on pairwise learning
  - It is not only good at feature learning (semi-supervised) but also be able to greatly integrate with supervised learning
    - Discriminative feature learning
    - Limited data
      - Small #data
      - Partial label

More information can be found at
https://cchsu.info