Unsupervised Domain Adaptation for Face Recognition in Unlabled Videos

Kihyuk Sohn Sifei Liu Guangyu Zhong Xiang Yu Ming-Hsuan Yang Manmohan Chandraker

IIE8557-01 동적계획법과 강화학습

경영과학연구실 전재현

Video Face Recognition

- 동영상에서 사람의 얼굴을 인식하는 것
- Image와 달리 labeling된 대규모 데이터셋이 존재하지 않음
- Frame 단위로 동영상을 잘랐을 때 still-image처럼 선명하지 않음
- Image dataset으로 학습시켰을 때 domain간의 간극이 존재

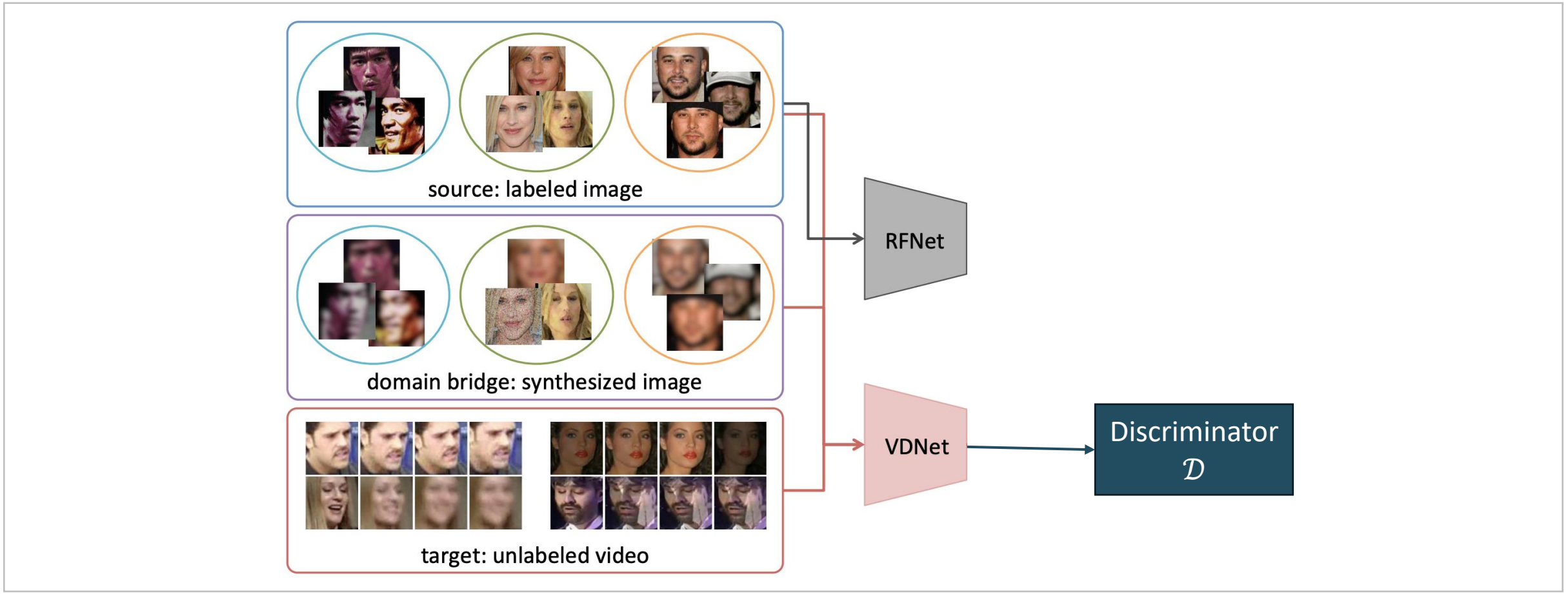
대규모 데이터셋이 부재한 상황에서의 Video Face Recognition을 연구

Key Idea

- Image Face Recognition with transfer learning
- Data augmentation(transformation)
- Adversarial learning

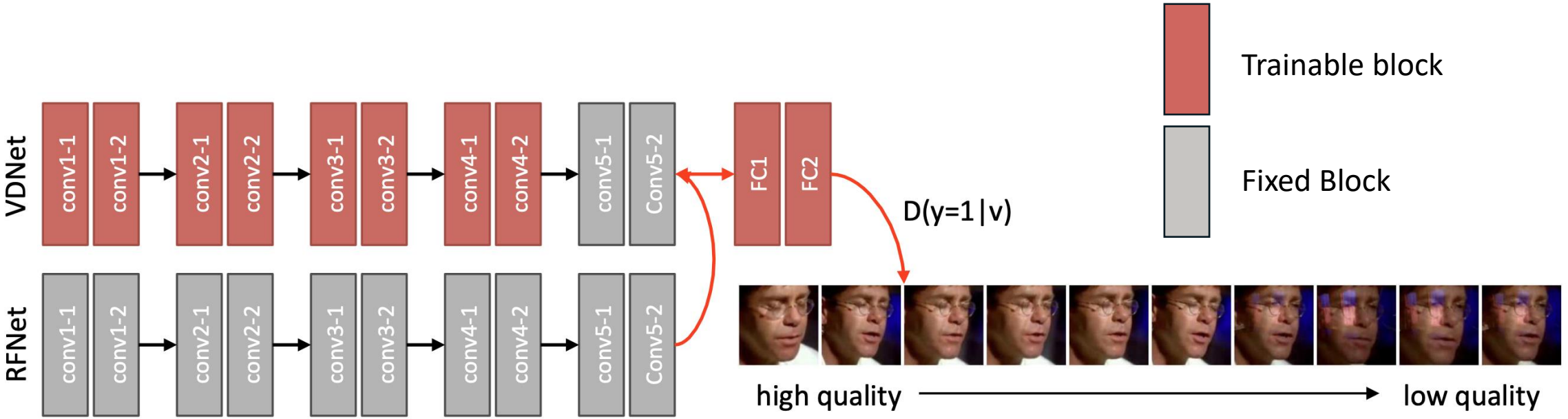
Overall Framework

- 대규모 image 데이터셋으로 학습한 RFNet을 활용하여 VDNet을 학습
- VDNet 학습에는 label된 image, synthesized image, unlabeled video가 함께 사용됨
- 3개의 서로 다른 범주의 data에서 비슷한 정보를 얻어내고자 함



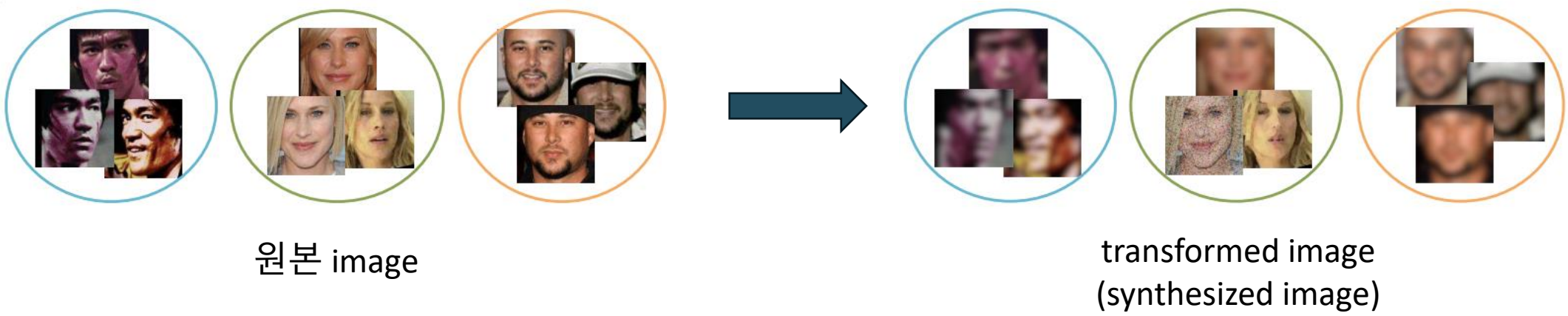
Distilling Knowledge

- Labeld web-face dataset으로 학습된 RFNet
- 일부 Layer를 고정시키고 RFNet으로부터 나온 feature와 VDNet으로부터의 feature 차이를 최소화하도록 나머지 Layer들을 학습



Data Augmentation(transformation)

- Video frame에서의 이미지는 StillI 이미지와 차이가 있음
- 따라서 motion blur나 scale variation 등의 방법을 통해 video와 비슷한 data 생성
- 원본 RFNet을 통해 추출된 원본 image의 feature와 VDNet을 통해 추출된 transformed image의 feature의 차이가 최소화되도록 하는 것이 목적

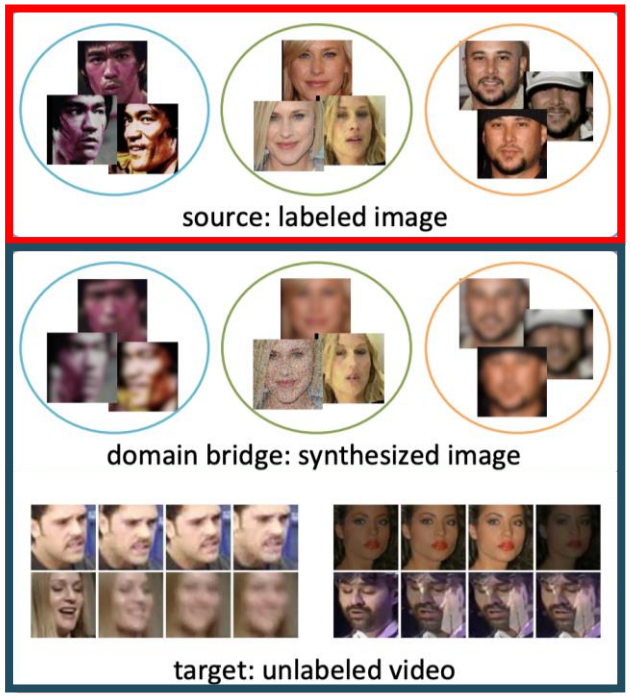


원본 image

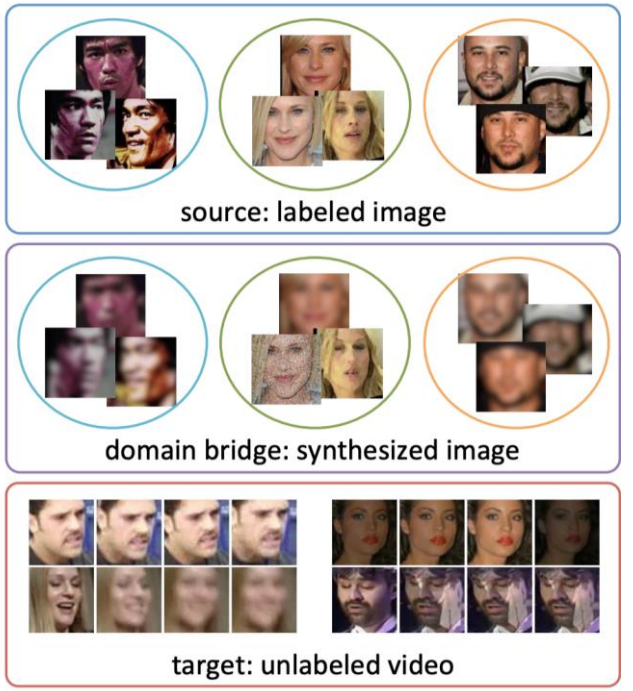
transformed image
(synthesized image)

Adversarial Learning

- 2-way : 원본 image vs. synthesized image, unlabeled video
- 3-way : 원본 image vs. synthesized image vs. unlabled video



2-way



3-way

Experiments

- YTF(youtube faces) dataset 사용
- Training에는 label이 없는 video를 사용
- 일정 frame 구간 내에 얼굴이 인식되는 정확도를 평가
- 정확도를 평가할 때는 label이 있는 video를 사용

Experiment on the YTF dataset

- 앞서 소개한 방법들에 대한 ablation study 진행
- Transfer learning, augmentation, adversarial learning을 모두 포함한 E, F 모델이 좋은 성능을 보임

Model	IC	FM	FR	Adv	fusion	1 (fr/vid)	5 (fr/vid)	20 (fr/vid)	50 (fr/vid)	all
baseline			–		–	91.12±0.318	93.17±0.371	93.62±0.430	93.74±0.443	93.78±0.498
					✓	–	93.30±0.362	93.72±0.428	93.80±0.444	93.94±0.493
A	✓	–	M/S	–	–	91.37±0.334	92.97±0.381	93.42±0.399	93.43±0.384	93.32±0.443
B	✓	✓	M/S	–	–	91.44±0.348	93.46±0.392	93.84±0.433	93.95±0.443	93.94±0.507
C	✓	✓	M/S/C	–	–	91.68±0.320	93.52±0.323	93.94±0.337	93.90±0.361	93.82±0.383
D	✓	✓	–	two-way	–	91.38±0.350	93.74±0.354	94.04±0.375	94.23±0.379	94.36±0.346
E	✓	✓	M/S/C	two-way	–	92.39±0.315	94.72±0.306	**95.13**±0.263	**95.13**±0.286	**95.22**±0.319
					✓	–	94.73±0.270	**95.14**±0.229	**95.13**±0.261	**95.16**±0.284
F	✓	✓	M/S/C	three-way	–	92.17±0.353	94.44±0.343	**94.90**±0.345	**94.98**±0.354	**95.00**±0.415
					✓	–	94.52±0.356	**95.01**±0.352	**95.15**±0.370	**95.38**±0.310

Experiment on the YTF dataset

- 다른 방법들과의 비교
- 비교적 구조가 복잡한 image-based 방식과 비교했을 때, 단순한 구조로 더 좋은 성능을 보여줌

Unsupervised DA		SOTA (image-based)	
baseline	93.78	DeepFace [32]	91.4
PCA	93.56	FaceNet [26]	95.12
CORAL [28]	94.50	CenterFace [36]	94.9
Ours (F)	**95.38**	CNN+AvePool [39]	95.20

Conclusions

- 라벨링된 대규모 video 데이터셋이 없다는 문제를 transfer learning, adversarial leraning 등을 활용하여 해결하려고 함
- 대규모 image 데이터셋으로부터 사전학습된 모델을 사용하여 비교적 단순한 구조로 높은 성능을 보여줌

Q & A