Check for updates

Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.867
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

A Deepfake Video Detection Considering Sequential Keyframes and An Estimation of Fake Area by Input Attribution

HaYoung Park[a], Young Han Lee[a], Sangkeun Lee[b], and Choongsang Cho[a]‡

Abstract

Recently, various generative models have been researched and are applied to visual content creation as their performance improves. On the other hand, deepfake frameworks with generative methods focused on faces including identity have been used to generate images and videos with different meanings than intended. Synthesizing pornographic videos or creating fake speaking videos with facial images, and then distributing them online is an increasingly serious problem. To protect people from the damages caused by deepfake technology used for illegal purposes, detecting the deepfake content and checking the synthesized part of the face are essential elements to minimize damages caused by computer vision technology. In this paper, we propose a deepfake video detection that analyzes face regions using sequential keyframe image pairs. Also, the synthesized part of the face is estimated by analyzing input attribution of the proposed structure with an explainable scheme. Our experiments show that detection performance improves as the number of images observed in the detection network at one time increases.

Keywords : Deepfake, Deepfake video detection, Sequential video keyframes, Fake region estimation

I. Introduction

Over the past few years, generative models such as generative adversarial networks (GANs) , variational autoencoders, and diffusion schemes have dramatically gathered interest in computer vision since they can be used in various content creations and building datasets for deep-learning[1-2]. The generated high-quality images and videos are difficult to distinguish whether original or fake with one's eyes. In addition, it is more difficult to distinguish due to the degradation of visual quality caused by video compression for transmission. A face includes a lot of identity information, and deepfake technology based on generative models is a useful method to create digital avatars through synthesis focused on the face[3-7]. However, face identity and intention can be changed through deepfake technology to create fake news and pornographic videos. Therefore, methods to detect deepfake videos and estimate the fake face areas

a) Korea Electronics Technology Institute
b) Smart Vision Global
‡ Corresponding Author : Choongsang Cho
E-mail: ideafisher@keti.re.kr
Tel: +82-31-739-7457
ORCID: https://orcid.org/0000-0001-5952-5491
※ This work was partially supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT)(No. 2021-0-00888, No. 2022-0-009984, No. 2022-0-00926)
· Manuscript October 4, 2023; Revised November 27, 2023; Accepted November 27, 2023.

Copyright © 2023 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

are essential techniques to reduce the damage of the wrong use of technology. In this paper, we propose a deep fake video detection approach that utilizes multiple feature extraction networks to analyze face region image pairs from sequential keyframes region under compressed video data[8]. Also, in order to estimate the fake face region and analyze the reliability of the operation of the proposed deepfake video detection, the input attribution to the results of the network is evaluated by an explainable method[9].

II. Related Work

To deceive human intention and identification in visual content, deepfake schemes have progressed to mimic and synthesize human faces. In deepfake generation schemes, the methods to synthesize face image in a given video into the other's face have been proposed in Deepfakes[3], FaceSwap[4], and FaceShifter[5]. A method referred to as Deepfakes requires one video pair at the training process for swapping operation. FaceSwap is conducted by the face-swapping application based on computer graphics. FaceShifter generates a face image using integrated features from the source and target, followed by the occlusion removal step. The methods mimicking the source face while preserving the identity of the target face have been proposed[6-7]. Face2Face[6] defines an energy function and mathematically optimizes it to combine features of source and target. The generated video focused on the mouth region in the deepfake dataset is especially conducted by Neural texture[7] method that uses the expression and identity characteristics of the source and target as input of neural render. In order to encourage the positive usages of generative methods, deepfake detection methods have been researched to prevent their negative purposes. A high performance detection scheme based on XceptionNet[10] and dataset are released in[8]. The augmentation method focused face structure and capsule networks are used in[11].

III. A Deepfake Video Detection Using Sequential Keyframe Features

We present a deepfake video detection framework that con-

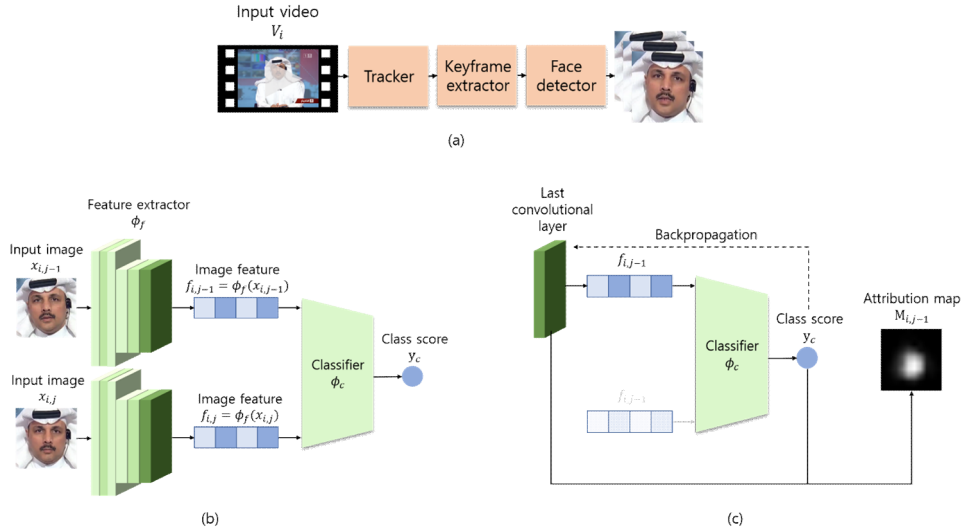

Fig. 1. Overall diagram of our proposed framework. (a) Extracting face images and their identification (id) from N keyframes of an input video, (b) Detecting deepfake via feature extractor and classifier where j is a keyframe index with range $2 \le j \le N$, (c) Computing input attribution map $M_{i,j-1}$.

siders the variation of face regions through sequential keyframe images. Fig. 1 shows the overall architecture of our proposed method that consists of (a) preprocessing step with a human tracker, a keyframe extractor, and a face detector, (b) deepfake detection step with feature extractor and deepfake classifier, and (c) estimating manipulated face region by obtaining input attribution map. The following sections provide details of the steps and describe the input attributions for the reliability assessment of our proposed scheme.

1. Preprocessing to Obtain Sequential Face Region

In [8], using cropped face images for a deepfake detection network provide much higher accuracy compared to using the entire images. On the other hand, utilizing all frames of a given video not only takes a high computational complexity for the decision of a video but also requires huge training and computation time. Therefore, we propose to use cropped face images of keyframes in a video, which can incorporate useful information compactly by removing less informative spatial backgrounds and temporal duplicated frames at the same time. Let $V=\{V_i\}_{i=1}^{M}$ denote the set of videos. During the preprocessing step, people present in a video V_i are traced by a trained tracker to maintain the identity of a person throughout sequential frames. Next, keyframe extraction to select the most representative frames is conducted. A simplified method from the complex keyframe extraction method[12] including computation of color histograms edge direction histograms, wavelet statistics, etc. extracts keyframes by computing the difference between successive frames and choosing N frames with the high variation score. Lastly, the set of face regions with the same identity in keyframes are obtained by a trained face detector to be used as inputs of a network that receives the sequential face images of the same human. The process is shown in Fig. 1 (a) and can be simply written as,

$$X_i^{id}=F_D\left(K_E\left(Track\left(V_i\right)\right)\right), 1\le i\le M \quad , \tag{1}$$

where V_i is $i-th$ video in the dataset, $X_i^{id}=\{x_{i,j}^{id}\}_{j=1}^{N}$ is a face image set of $i-th$ video with an identity id. F_D, K_E, and $Track$ indicate a face detector, a keyframe extractor, and a tracker, respectively. In the case of only a single person present in the video, $x_{i,j}^{id}$ can be simplified as $x_{i,j}$. In the following sections, we assume one person per video and omit id for a better understanding. Then the set of extracted face images of $i-th$ video can be represented as: $X_i=\{x_{i,j}\}_{j=1}^{N}$.

2. Deepfake detection with sequential face image pairs

Deepfake generation methods produce highly realistic fake images indistinguishable by a single image, on the other hand, their movement in a video is often unnatural and detected as a synthetic result. Therefore, we propose a deepfake detection framework that receives multiple face images of a single human as input to consider the movement information in a video. Also, under the input conditions, configuring multiple face images of a single human extracted from keyframes becomes one of the crucial points to avoid almost identical face image pairs in a video.

As shown in Fig. 1 (b), the feature extractor ϕ_f extracts feature vectors from the sequential face image of the identical person, and the classifier ϕ_c estimates the probability of fakeness. Let $X=\{x_{i,j}\mid x_{i,j}\in X_i, 1\le i\le M, 1\le j\le N\}$ denote the set of face images from keyframes through the preprocessing steps, then, obtaining the fakeness probability can be written as:

$$p_{i,j}=\mathrm{softmax}\left(\phi_c\left(f_{i,j-1},f_{i,j}\right)\right), f_{i,j}=\phi_f\left(x_{i,j}\right), \tag{2}$$

where $p_{i,j}$ is the output fakeness probability of the given input set, $(x_{i,j-1}, x_{i,j})$, and $f_{i,j}$ is an extracted feature by ϕ_f . The fakeness decision of the paired keyframes is conducted by comparing the probability with the given threshold as $c_{i,j} = \psi_t(p_{i,j}; \tau_{img})$, where $c_{i,j}$ is a predicted class, τ_{img} is a threshold for an paired keyframe, and the decision function, ψ_t, is defined as

$$\psi_t(z;\tau) = \begin{cases} 0 & if\ z < \tau, \\ 1 & \text{otherwise}, \end{cases} \quad (3)$$

where 0 and 1 indicate the real and fake, respectively. Deciding whether a video is manipulated or not is computed by measuring the average probability over all paired keyframes as

$$c_i = \psi_t\left(\frac{1}{N}\sum_{j=1}^{N} c_{i,j}; \tau_{vid}\right), \quad (4)$$

where c_i is a predicted class of a video and τ_{vid} is a threshold for a video.

3. Estimation of fake face region by input attribution analysis on the network

Since the results of the deepfake discrimination are highly related to the fake region, the input contribution of the results is highly relevant to the synthesized face part. Therefore, to estimate the fake face area, input attribution in the deepfake decision model is measured by computing the relationship between the input and the output through a gradient-based class activation mapping (CAM)[9]. The input attribution of the proposed network is obtained by using the CAM as,

$$M_{i,j} = \psi_m\left(x_{i,j}, \phi_c\left(\phi_f(\cdot)\right)\right), \quad (5)$$

where ψ_m denotes the CAM method and $M_{i,j}$ is an input attribution represented by its heatmap.

IV. Experiments

In this section, we show that the proposed sequential keyframe based deepfake detection network achieves better performance compared to the method with a single input image.

1. Implementation details

1) Dataset: The deepfake data set, FaceForensics++[8], consisted of 6, 000 videos including 1, 000 real videos from Youtube and 5, 000 generated videos by Deepfakes[3], Face2Face[6], Faceshifter[5], FaceSwap[4], and NeuralTextures[7]. For our experiments, the H.264-compressed videos with a compression rate factor of 23 are used in the dataset. For the training, validation, and test sets, we split the videos into training (80%), validation (10%), and test (10%) sets. Also, a maximum keyframe from a video is set 15 for the preprocessing of videos in the dataset. To evaluate the performance of the proposed structure, the detection accuracy of image-wise and video-wise was respectively measured to recognize the generated image and video by deepfake methods. Thresholds τ_{img} and τ_{vid} to decide the fakeness are set to 0.5. In addition, we employ a logloss method to evaluate the quality of predicted probabilities.

2) Training setup: We implemented the proposed network using EfficientNet, EfficientNetV2, and Swin Transformer as a backbone network. To fairly compare with previous research experiments of[8], we also report the results when XceptionNet is used as a backbone network under identical settings with the proposed method. The SGD (Stochastic Gradient Descent) optimizer with a learning rate of 0.01 was used for the network based on EfficientNets and AdamW optimizer with a learning rate of 0.0001.

2. Deepfake detection results

Under various backbone networks, the proposed method

Table 1. Comparison of the conventional scheme and our proposed method with several backbone networks, image-wise deepfake detection accuracy and log-loss. Multiple (2) means two face images of a human from keyframes as the inputs

	Backbone	Number of inputs	Accuracy	Log Loss
Conventional	XceptionNet [8]	single	94.036	0.371
	XceptionNet [8]	Multiple (2)	94.320	0.346
Conventional method with backbones	EfficientNet-b3	Single	94.417	0.392
	EfficientNet-b4	Single	94.464	0.179
	EfficientNetV2-s	Single	95.333	0.200
	EfficientNetV2-m	Single	95.036	0.276
	SwinT-tiny	Single	94.369	0.199
	SwinT-small	Single	93.952	0.263
Ours	EfficientNet-b3	Multiple (2)	96.120	0.273
	EfficientNet-b4	Multiple (2)	95.801	0.169
	EfficientNetV2-s	Multiple (2)	95.992	0.243
	EfficientNetV2-m	Multiple (2)	95.763	0.176
	SwinT-tiny	Multiple (2)	94.550	0.218
	SwinT-small	Multiple (2)	95.111	0.208
	EfficientNetV2-s	Multiple (3)	96.342	0.245
	SwinT-tiny	Multiple (3)	95.517	0.169

was compared with a single input condition like the conventional method[13]. As shown in Table 1, the proposed method using two sequential keyframes, multiple (2), achieves higher accuracy compared with the detection method using an image under backbone networks. In addition, the proposed scheme with three sequential keyframes, multiple (3), shows better performance than the proposed method with two sequential keyframes. Table 2 shows video-wise deepfake detection accuracy according to fake generation methods that include Deepfakes (DF), Face2Face (F2F), FaceShifter (FSH), FaceSwap (FSW), and NeuralTextures (NT). The accuracy of real videos and generated videos by NeuralTextures is lower than the results of the other methods. Because the NeuralTexures method only focuses on changing the mouth region, real videos are confused with them.

Table 2. The accuracy of detecting deepfake videos using the proposed method with EfficientNetV2-m backbone

RL	DF	F2F	FSH	FSW	NT	total
95.19	100.0	98.88	100.0	100.0	93.75	97.88

3. Fake region estimation by the explanation of a detection network

To estimate the fake face region by an explanation of the proposed network, we employ GradCAM++[9] to obtain input attribution maps that reflect the influence of inputs on results. As shown in Fig. 2, the overlaid images of input and their maps show the input attribution to results in the first row. The red and blue colors indicate the high and low attribution scores, respectively. Also, to analyze the influence characteristic of the input according to the deepfake generation methods, the average attribution map according to the generation methods is calculated using all test videos, as shown in the second row in Fig. 2. The proposed detection network well focuses on the face that has a high contribution score to the decision. Furthermore, statistical analysis using the input attribution mean shows a shape related to the characteristics of the deepfake technique, and the fake region is concentrated according to the deepfake methods. In more detail, the attribution maps of face-swapping methods (Deepfakes, FaceShifter, and FaceSwap)

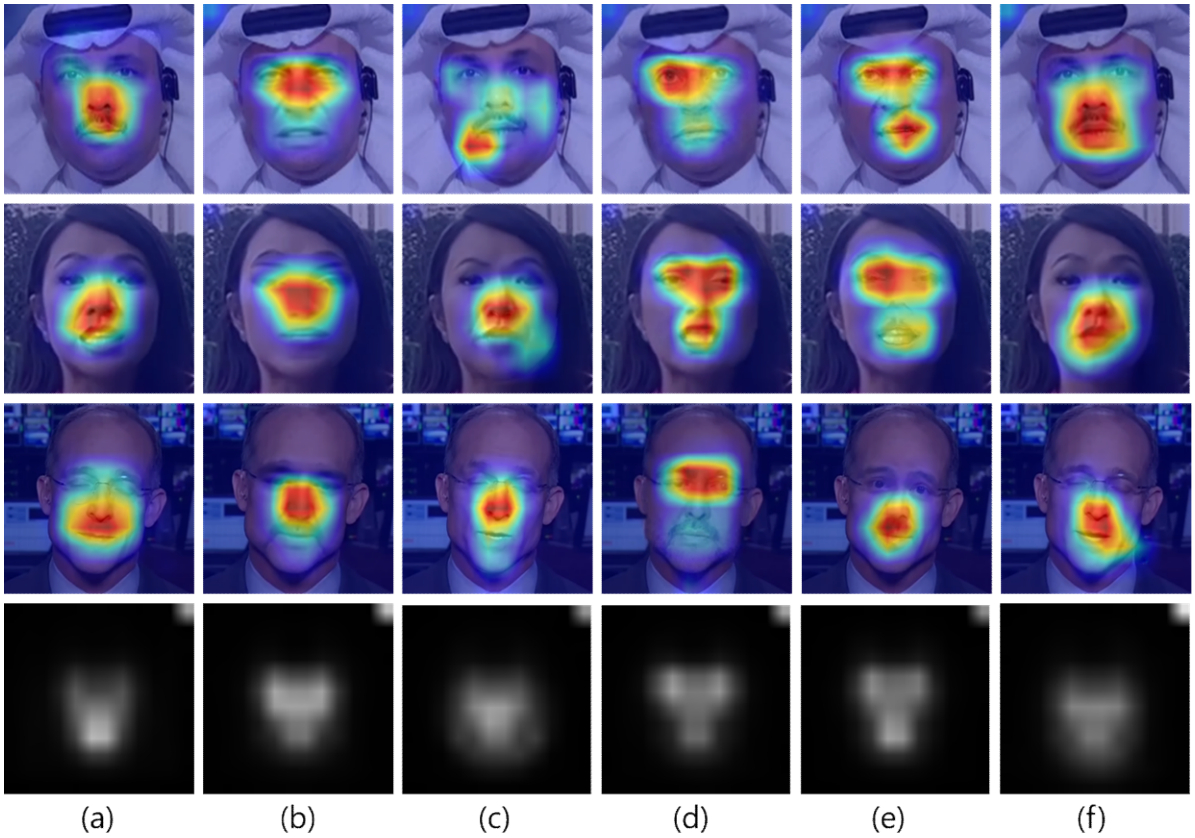

Fig. 2. The images in the top three rows, from left to right, show a face image overlaid with an attribution map of a real image (a), a synthesized face image by Deepfakes (b), Face2Face (c), FaceShifter (d), FaceSwap (e), and NeuralTextures (f), respectively. The images in the last row are the accumulated input attribution maps of test videos, which provide an explanation of the proposed network based on the generation schemes.

show high attention scores on crucial face components; eyes, nose, and mouth region. Since these methods change the identity of a face by replacing the entire face region with that of another face. In the reenactment schemes (Face2Face and NeuralTextures), attribution maps on the face components have lower scores than the swapping methods, on the other hand, the broad area of the face contributes to the attribution map of input. The attention region of Face2Face is wider than that of NeuralTexture since Face2Face modifies all face components for the reenactment, while NeuralTextures only modifies the mouth region.

V. Conclusions

In this paper, a deepfake detection framework using face image pairs of a single human from keyframes in a video and an estimation of fake face region with input attribution were presented. Also, the fake face region by the deepfake methods is estimated by the input attribution extracted from an explainable scheme of the network. The experimental results showed that the proposed structure improved their performances as the number of images observed in the detection network at one time increased. The averaged attribution maps according to the deepfake methods showed clearly different maps depending on the deep-

fake methods and the characteristic of estimated region by input attribution is highly relevant to the properties of deepfake generative methods. Therefore, we believe that the proposed deepfake detection and analysis of its results can be a useful tool to reduce the damage by illegally used deepfake technology.

References

[1] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," in Proceeding of Neural Information Processing Systems (NeurIPS), vol. 32, 2018.
https://proceedings.neurips.cc/paper_files/paper/2019/file/370bfb31abd222b582245b977ea5f25a-Paper.pdf
doi: https://dl.acm.org/doi/10.5555/3454287.3454738

[2] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially adaptive normalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2337 – 2346, 2019.
https://openaccess.thecvf.com/content_CVPR_2019/papers/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.pdf
doi: https://doi.org/10.1109/cvpr.2019.00244

[3] "Deepfakes github," https://github.com/deepfakes/faceswap. Accessed: 2018-10-29.

[4] "Faceswap github," https://github.com/MarekKowalski/FaceSwap/. Accessed: 2018-10-29.

[5] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5074 – 5083, 2020.
https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Advancing_High_Fidelity_Identity_Swapping_for_Forgery_Detection_CVPR_2020_paper.pdf
doi: https://doi.org/10.1109/cvpr42600.2020.00512

[6] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387 – 2395, 2016.
https://openaccess.thecvf.com/content_cvpr_2016/papers/Thies_Face2Face_Real-Time_Face_CVPR_2016_paper.pdf
doi: https://doi.org/10.1109/cvpr.2016.262

[7] J. Thies, M. Zollhofer, and M. Nießner, "Deferred neural rendering: ¨ Image synthesis using neural textures," ACM Transactions on Graphics (TOG), vol. 38, no. 4, pp. 1 – 12, 2019.
https://dl.acm.org/doi/pdf/10.1145/3306346.3323035
doi: https://doi.org/10.1145/3306346.3323035

[8] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. ¨ Nießner, "FaceForensics++: Learning to detect manipulated facial images," in International Conference on Computer Vision (ICCV), 2019.
https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf
doi: https://doi.org/10.1109/iccv.2019.00009

[9] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839 – 847, 2018.
https://www.computer.org/csdl/proceedings-article/wacv/2018/488601a839/12OmNs5rkO4
doi: https://doi.org/10.1109/wacv.2018.00097

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251 – 1258, 2017
https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf
doi: https://doi.org/10.1109/cvpr.2017.195

[11] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307 – 2311, 2019.
https://ieeexplore.ieee.org/document/8682602
doi: https://doi.org/10.1109/icassp.2019.8682602

[12] G. Ciocca and R. Schettini, "Dynamic key-frame extraction for video summarization," SPIE Internet Imageing VI, vol. 5670, pp. 137-142.
https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5670/0000/Dynamic-key-frame-extraction-for-video-summarization/10.1117/12.586777.short?SSO=1
doi: https://doi.org/10.1117/12.586777

[13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in Proceedings of Machine Learning Research (PMLR), pp. 6105 – 6114, 2019.
https://proceedings.mlr.press/v97/tan19a/tan19a.pdf

Introduction Authors

Hayoung Park

- Hayoung Park received both his B.S. and M.S. degrees in electronics engineering from Hanyang University, Gyeonggi, Korea. Since 2021, he has been a researcher with the department of intelligent information research, Korea Technology Institute (KETI).
- Research interests : supervised, semi-supervised, weakly supervised visual intelligence, as well as its domain adaptive technology.

Young Han Lee

- Young Han Lee received a B.S. degree in electronics engineering from Gwangwoon University, South Korea in 2005, and an M.S. and Ph.D. degrees in electrical engineering and computer science from Gwangju Institute of Science and Technology (GIST), in 2007 and 2011, respectively. From 2011 to 2014, he was a Senior Researcher with LG Advanced Research Institute, South Korea. Since 2015, he has been an principal researcher with the department of intelligent information research, Korea Technology Institute (KETI).
- ORCID : https://orcid.org/0000-0001-8200-2867
- Research interests : speech and audio signal generation and multimodal processing.

Sangkeun Lee

- Sangkeun Lee (SM'12) received the B.S. and M.S. degrees in electronic engineering from Chung-Ang University, Seoul, South Korea, in 1996 and 1999, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2003.
  He was a Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul. From 2003 to 2008, he was a Staff Research Engineer with the Digital Media Solution Lab, Samsung Information and Systems America, Irvine, CA, USA, where he was involved in the development of video processing and enhancement algorithms (DNIe) for Samsung's HDTV. He is developing a PBM-based electronic medicine for the treatment of eye diseases at a startup.
- ORCID : https://orcid.org/0000-0001-6589-6774
- Research interests : Digital video/image processing, image understanding, computer vision, deep learning, AI-based medical device, and CMOS image sensors.

Choongsang Cho

- Choongsang Cho received the B.S. degree in electronics engineering from Suwon University, Gyeonggi, Korea, and the M.S. degree from the Department of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Korea. He received the Ph.D. degree with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Seoul, Korea, in 2016 . Since 2008, he has been an principal researcher with the department of intelligent information research, Korea Technology Institute (KETI).
- ORCID : https://orcid.org/0000-0001-5952-5491
- Research interests : self-supervised, unsupervised and explainable visual intelligence technology.