International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Emerging Trends and Innovations in Web-Based Applications and Technologies
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Meta-Learning Method for Few-Shot Face Forgery Segmentation and Classification

Govind Raut[1], Prof. Harshita[2], Chandrakant Kottalwar[3], Prof. Anupam Chaube[4]

[1,2,3,4]Department of Science and Technology,
[1,2,3]G H Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India
[4]G H Raisoni College of Engineering and Management, Nagpur, Maharashtra, India

ABSTRACT

Detection of Forgeries in Images: A Survey Abstract: While the technology to detect forgeries in images is able to detect images even at complex images, this well is only limited to known forgery methods. It trains neural networks from large amounts of original and corresponding forged images created with known techniques. But it fails to process unseen forgery techniques. One such proposed solution to this problem, recently, is to employ a hand-crafted generator of forged images to generate a series of fake images and feed them to the neural network for training However, the aforementioned approach has certain limitations detecting performance in situations where the hand-craft generator has not taken into consideration invisible forging processes. In this study, we use a meta-learning approach to create a highly adaptive detector for detecting novel forging techniques, overcoming the drawbacks of current approaches. By employing meta-learning approaches to train a forged image detector, the suggested method allows the detector to be fine-tuned using a small number of fresh forged examples. In order to detect forged images with comparable characteristics, the suggested method inputs a limited number of the forged images to the detector and allows the detector to modify its weights based on the statistical properties of the input forged photos. With IoU gains ranging from 35.4% to 127.2%, the suggested approach significantly improves forgery method detection. These findings illustrate that the suggested approach outperforms the state-of-the-art techniques in the majority of situations and greatly enhances detection performance with a relatively small number of samples.

1. INTRODUCTION

The development of deep learning in recent years has tremendously improved the problem of falsified facial photos as a security threat. Deep learning has also been used for detection in order to conduct forensic analysis on these kinds of falsified photos. Presently available technology for identifying forged photos does a good job of identifying established forging techniques. This system trains neural networks, which serve as detectors for learning the features of forged images, using a huge number of original and related forged images produced using known forged techniques. Nevertheless, when these techniques come against untrained forged methods, their detection efficiency significantly declines. Recently, an innovative approach has been put out to overcome this problem [1]. Using a parameterizable forged image generator, this technique generates a variety of forged images, which are subsequently used to train a neural network. The forged image generator is trained using the immaculate image, as shown in Figure 1. Forging technique A and forging method B are two of the mechanisms used by the forged image generator/synthesizer G to produce forged images. The forged picture detector is then trained using these artificially created forged images. Given that the generator of faked images was created by

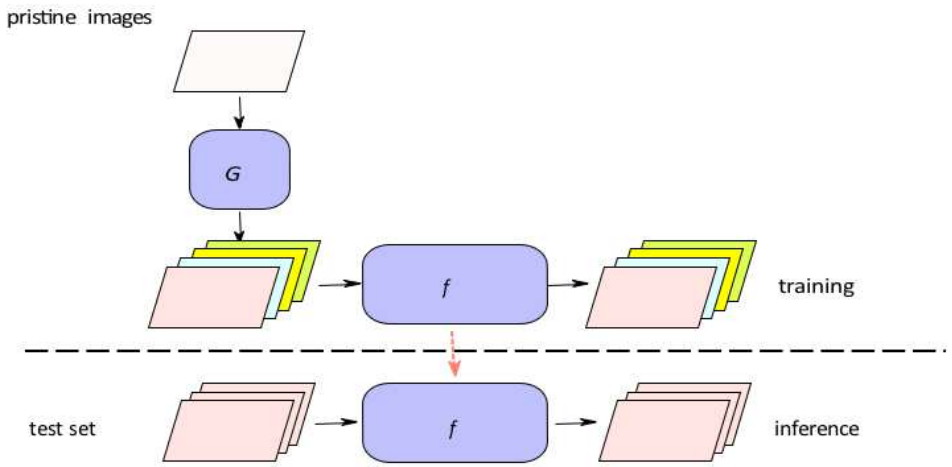


Figure 1. shows earlier techniques for synthesizing fictitious samples. Using a fake image synthesizer G, the pristine photographs are used to create fake images based on color jitter, scaling, sharpening, and translation. The employment of training weights during the inference phase is indicated by the orange dashed line.

We provide a meta-learning strategy to train a detector that is highly effective at spotting novel forging processes in order to solve the problem with the previously discussed approaches. Our approach to detecting forged photos is based on training a forged image detector with a small number of new, forged examples, using meta-learning techniques. Training the forged picture detector is intended to allow it to accept a limited quantity of fresh, forged image samples and modify their weights to find fake photos that have statistical characteristics similar to the limited sample of fake photos that were supplied.

2. Related Work

There are several techniques for making false faces in the literature. We will particularly present a few techniques that are pertinent to our study in this paper. The NeuralTextures approach, put forth by Thies et al., is one such technique. Using a rendering network and a unique algorithm, this technique enhances the quality of a computer-generated texture to produce a realistic reenactment. Another technique, called the Face2Face facial reenactment system, converts 2D facial points into 3D models from source video streams and combines the 3D models' modified faces with other facial traits. FaceSwap, a third technique, is a method based on computer graphics that creates altered facial features by mapping the facial landmarks of source faces onto a 3D template model. The DeepFakes methodology, which is based on deep learning, is an additional way to create synthetic faces. In order to create the target fake faces, this technique first extracts faces from the original images, then uses a trained encoder and decoder for the source faces. This method, which has drawn a lot of interest lately because of its potential for malicious use in creating convincing false films, has been demonstrated to create incredibly lifelike fake faces.

Automated techniques for identifying phony faces have been developed using recent developments in deep learning. Several CNN-based methods for forgery detection have been put forth in the literature. For instance, Rössler et al. proposed the use of a CNN-based model, namely XceptionNet, to address the forgery detection task as a binary classification problem. Nguyen et al.

The authors propose a new method for detecting and segmenting manipulated facial images, which they see as both a classification and a segmentation problem. They use auto-encoders and a specialized Y-shaped decoder to identify and mark the fake regions of an image.

3. The Proposed Scheme

3.1. Architecture of the Model

for $i \leftarrow 1$ **to** N **do**

Sample k images and their ground truth from support set

$S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_k, \mathbf{y}_k)\}$

Sample q images and their ground truth from query set

$Q = \{(\mathbf{x}'_1, \mathbf{y}'_1), (\mathbf{x}'_2, \mathbf{y}'_2), \ldots, (\mathbf{x}'_q, \mathbf{y}'_q)\}$

$\theta_i \leftarrow \varphi$. set initialization weight for each task

while not done **do** . gradient descent for optimizing θ

Evaluate $\nabla_{\theta i} L(f_{\theta i}(\mathbf{x}_j), \mathbf{y}_j)$ for $1 \le j \le k$ 9 Update $\theta_i \leftarrow \theta_i - \zeta \nabla_{\theta i} L(f_{\theta i}(\mathbf{x}_j))$ for $1 \le j \le k$

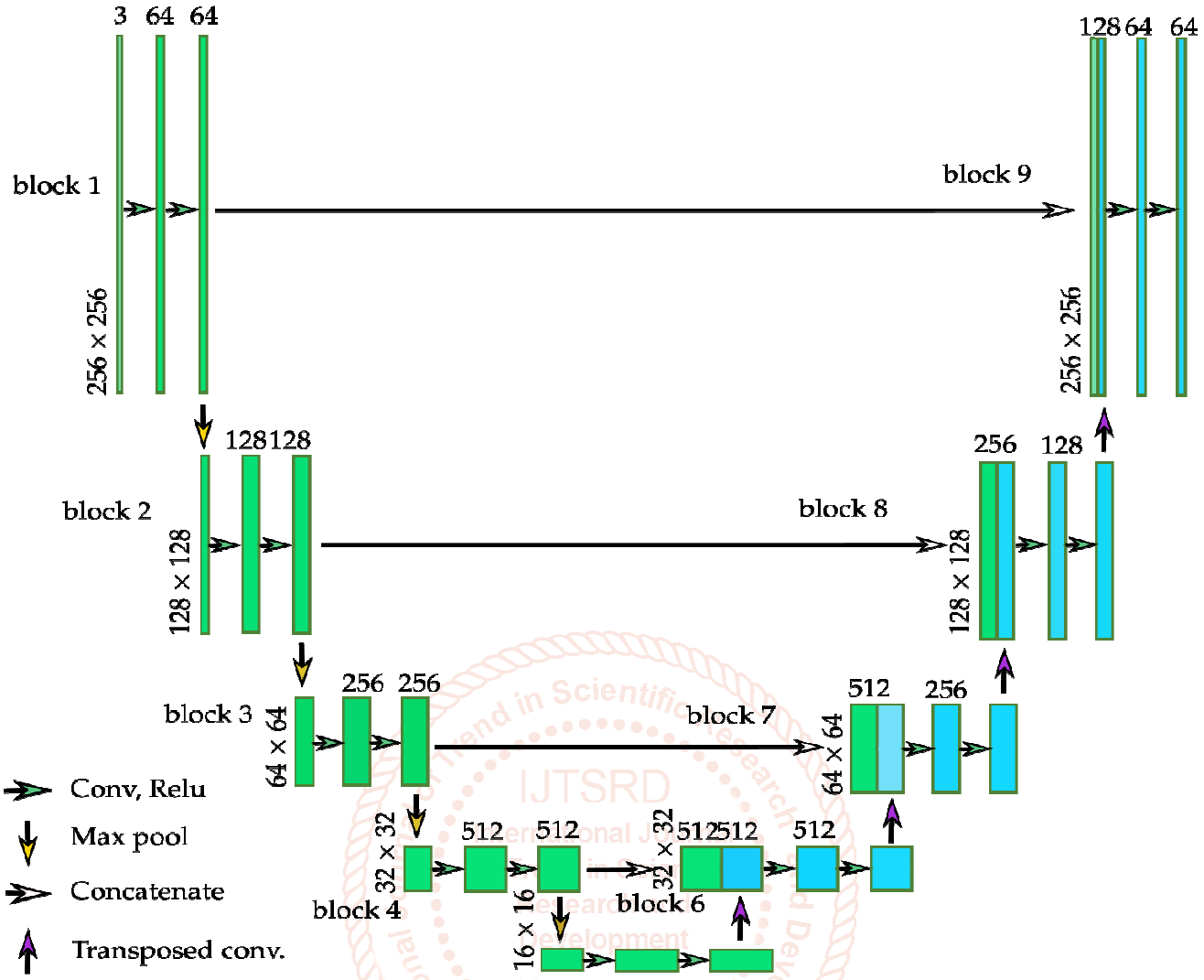
$\ell_i \leftarrow L(f_{\theta_i}(\mathbf{x}'_j), \mathbf{y}'_j)$ for $1 \le j \le q$. count loss using query set

Update $\phi \leftarrow \phi - \eta \nabla_{\phi} \frac{1}{N} \sum_{i=1}^{N} \ell_i$ *end*

To create the segmentation of projected manipulated regions, we employ the U-Net architecture, a variation of a fully convolutional network, to accept an input image and estimate the chance of each pixel being fake. The U-Net is made up of convolutional blocks and transposed convolutional blocks. Figure 3 depicts the detailed design of the U-Net utilized in the suggested technique. The U-Net receives the RGB-only input image x and produces the expected mask mˆ with only one channel. The input photos are enlarged to 256 × 256, and the pixel values are normalized using a mean of 0.5 and a standard deviation of 0.5 for each R, G, and B.channel. That is, the output of each channel equals the input minus the channel's mean divided by its standard deviation. The normalized photos (256 × 256 × 3) are then transmitted to U-Net as input. The U-Net convolutional block sequence extracts the fakeness feature, while the anticipated mask mˆ is synthesized using the concatenated sequence of the transposed convolutional block.

The training set includes both the altered image and its altered area. This altered area, also known as a mask, is used to identify which pixels are being modified during the forging process. Thus, the mask of a changed image serves as the ground truth for the forged area prediction issue (also known as the segmentation problem). Because the photos in the dataset include not just facial features but also significant sections of background, we utilize the mask to detect the location of the face in the image, and the cropped face is chopped from the surrounding area. If face detection is used to locate the face in the image and crop the face portion of the image, there will

The failure occurred because the face detection system was unable to recognize the falsified image. Cropped photos are shrunk to 256 × 256 and normalized with a mean and standard deviation of 0.5. The original image's rectangle area centered on the face is cropped and used as the input image for U-Net.



3.2. The Meta-Learning Approach

A fake segmentation task trains the algorithm to predict altered pixels in input photos using a training set created using a specific falsified approach. The model is trained using N fake segmentation tasks that can be easily changed to previously unseen counterfeit methods. See Algorithm 1's for loop, lines 3-10. The technique aims to identify parameters φ that may be taught with a small number of data to recognize fraudulent photos of previously unknown methods. When training on task i, the inner loop of the meta-learning in lines 7-9 employs gradient descent to update the model's weights, θi, with one or a few iterations using the support set, S = {(x1, y1), (x2, y2),..., (xk, yk)}. $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_k, \mathbf{y}_k)\}$, This is randomly selected from the task's dataset. Following the completion of the loop in lines 7-9, line 10 calculates the loss on the query set Q,..., using the current

weights (θi). After training N tasks in lines 4-10, line 11 updates parameters φ using gradient descent by taking the average of all loss 'i from the tasks. This technique is continued until φ meets the loss requirements. Meta-training is accomplished at this time, and the process moves on to the few-shot learning stage.

In the few-shot learning stage, K false photos from an undetected forging method U are used to fine-tune the optimal model from the previous stage using standard gradient descent. After training is completed, we test the model using fabricated photos created using forging method U.

To simplify, we construct the loss function as two one-dimensional vectors: yˆ = yˆ1, yˆ2,..., yˆn, and y = y1, y2,..., yn. This definition is easily applicable to higher-dimensional arrays.

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (-y_i \times \log(\sigma(\hat{y}_i)) - (1 - y_i) \times \log(1 - \sigma(\hat{y}_i))) \quad (1)$$

where $\hat{\mathbf{y}}$ is the prediction, and $\mathbf{y}$ is the ground truth, e.g., the output of U-Net $f_\varphi(\mathbf{x}) = \hat{\mathbf{y}}$.

where yˆ is the prediction and y is the ground truth. For example, the output of U-Net is fφ(x) = yˆ. The logistic sigmoid function is defined as σ(x) = 1+1ex. The function σ(·) is used to add nonlinearity to the output of neurons. This function also has the added benefit of restricting the neurons' output range between 1 and 0, allowing the output to be interpreted as a probability of confidence in predicting the fakeness of pixels. Algorithm 1 relies heavily on the fact that σ(·) is differentiable. The derivative is then used to update the weights of the U-Net fφ, with the goal of minimizing the error between the the expected and actual outputs. The primary goal of Formula (1) is to measure the difference between the segmentation's expected and true probability distributions (binary pixel categorization). The loss function is calculated as the negative loglikelihood of the true class given the predicted probability distribution of segmentation. Intuitively, the loss function determines how well the

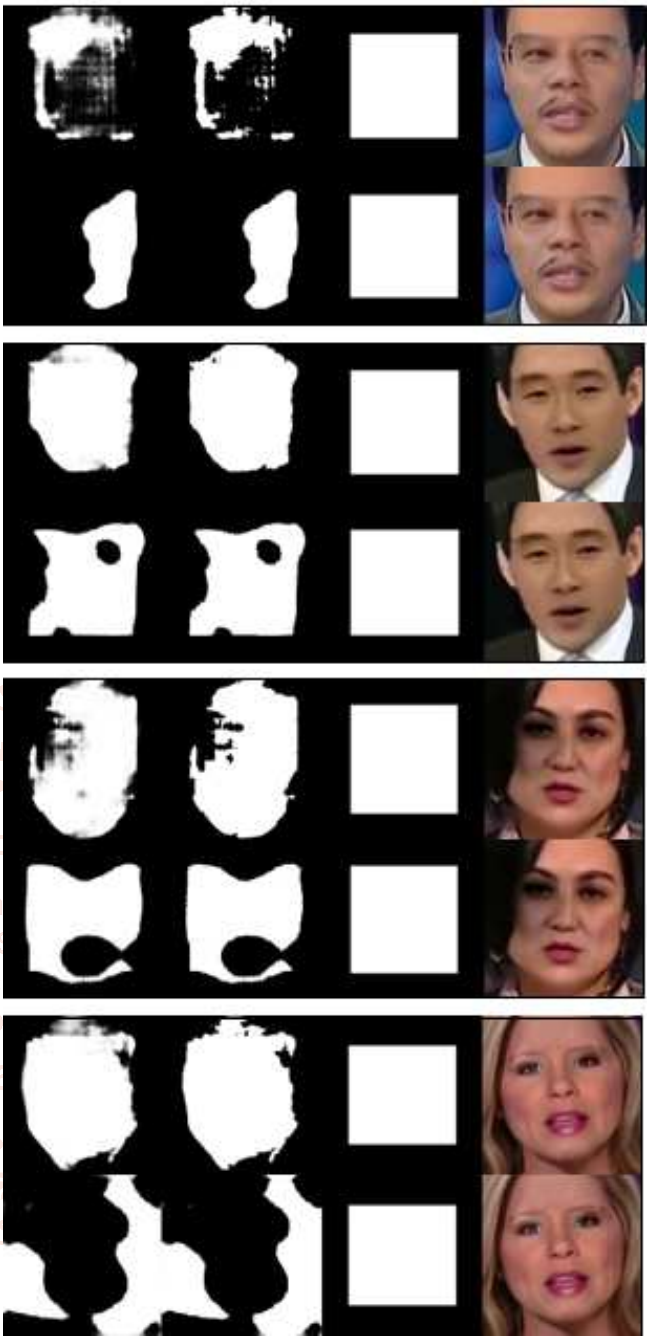
expected distribution matches the actual segmentation distribution. If the predicted distribution diverges much from the true distribution of segmentation, the loss will be severe. indicating a high degree of ambiguity or chaos in the prediction. If the expected distribution of segmentation closely matches the true distribution of segmentation, the loss will be modest.

4. Experiment and Comparison

4.1. Experimental Design and Data Collection

Images from the FaceForensics++ dataset, edited with the DeepFakes, Face2Face, FaceSwap, and NeuralTextures methods, were used in the trials to confirm the success of the proposed strategy. In this project, we blended authentic photographs with manufactured images using four different techniques: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The resulting images were then submitted to the proposed detector. In each trial, there is a 50/50 split between bogus and actual images. This mixed group of photos is fed into the detector we recommended to identify fabricated parts and determine whether the image is genuine. We may compare our findings to those of other relevant research that have used the same dataset, FaceForensics++, to evaluate the detection capabilities of Their algorithms and ability to detect phony pictures. We used images from the FaceForensics++ dataset in C23 format, compressed with H264 and a constant rate quantization setting of 23. C23 photos are used to simulate real-world conditions in which compression or other variables might reduce the quality of edited photographs. A high compression ratio, such as c40, will render the image highly fuzzy. A blurry image like this cannot be used in everyday situations, even though it is difficult to discern if it is genuine. The developers of the FaceForensics++ dataset acquired the 1000 flawless films from YouTube. The FaceForensics++ dataset is made up of 1,000 flawless videos.

Figure 1. Randomly selected DeepFakes photos and their anticipated segmentation results with one-shot fine-tuning are shown, with the top sub-row exhibiting the results acquired using the proposed method and the bottom sub-row displaying the results obtained without it. DeepFakes' altered images appear on the far right of each sub-row, followed by the ground truth of the altered region(mask), the binary predicted output, and the grey-scale predicted output in that order, from right to left.



DeepFakes left.

Figure 2. shows four rows of randomly picked DeepFakes photos and the anticipated results for the fake region using one-shot fine-tuning. Each row's top sub-row is the result of the proposed approach, while the bottom sub-row is the consequence of not employing the proposed method.

Figure 3. Comparison of AUC between random initial weights (without the proposed method) and meta-learning of detecting images altered by Face2Face manipulation methods. The x-axis is the size of the fine-tuned training set and the y-axis is the value of AUC.

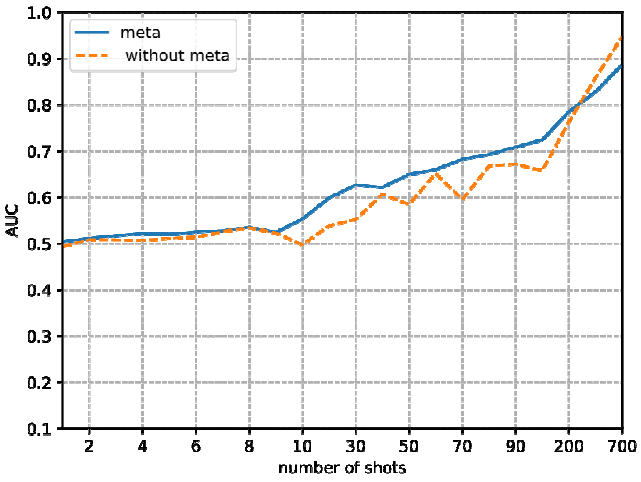


Figure 4. AUC comparison between random initial weights (without the suggested method) and meta-learning for picture detection using FaceSwap alteration methods. The y-axis represents the AUC value, while the x-axis represents the size of the fine-tuned training set.

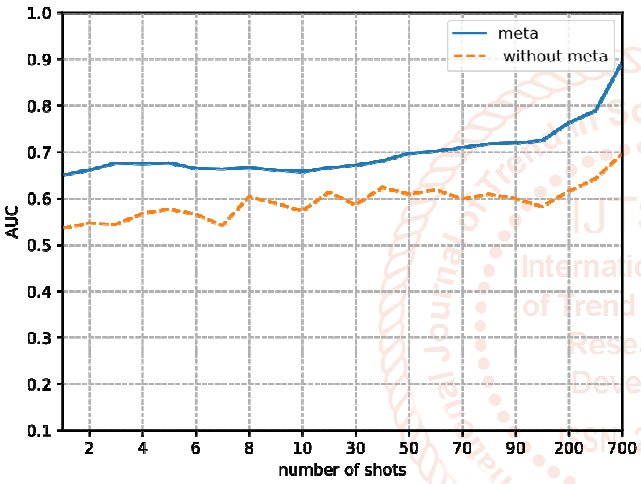


Figure 5. AUC comparison between random initial weights (without the suggested method) and meta-learning for image detection using NeuralTextures manipulation methods. The y-axis represents the AUC value, while the x-axis represents the size of the fine-tuned training set.

The most similar study to ours in published literature is [10], because our pioneering work seeks to identify forging zones and determine whether a certain image is fabricated using limited samples. Their work only examined two training sets and presented experimental results on the pixel-wise accuracy of forgery region detection, despite the fact that it detects forgeries and determines whether an input image is counterfeit. In Section 4.2, we explained why pixel-wise precision is not an appropriate criterion for detecting forgeries regions. However, the pixel-wise accuracy and IoU metrics employed in this study's comprehensive findings on counterfeit zone identification can also be used as a standard to measure the performance of future research endeavors. Table 2 compares the zero-shot outcomes of several detection strategies used in [10,38] to the suggested method. Table 2 shows the first two techniques provided by Cozzolino et al. [38]: FT_Res and FT. Nguyen et al. propose four techniques: deeper_FT, MT_old, no_recon, and MT_new. According to Table 2, the suggested methodology is the best at evaluating if the unseen approaches DeepFakes, Face2Face, and NeuralTextures are fake, while MT_Old is the best at detecting the unseen method FaceSwap.

5. Conclusion

Instead of developing a false picture detector with a big training dataset encompassing a variety of forgery techniques, this study used meta-learning to train a neural network capable of recognizing phony photos produced by multiple undetectable forgery strategies with a small number of samples. The suggested technique emphasizes the usage of data from a small number of samples in order to rapidly update the false detector. Despite the limited sample size, the experimental findings suggest that the proposed approach can greatly increase performance metrics such as AUC, accuracy, and IoU. This demonstrates that this strategy is worth examining further. Improving feature extraction from a small number of samples and broadening the range of possible techniques (training)Tasks are prospective future areas. This paper demonstrates that by employing the meta-learning paradigm, we can train a system to detect emerging counterfeit tactics from small sample numbers. As a result, new counterfeit tactics can be discovered with a minimal number of samples. As a result, the detector's response time can be reduced when competing with forgers. One of the disadvantages of our strategy is that it requires only a modest amount of training data. However, given the unavailability of a method that can detect every new forging technique without further training, obtaining a limited number of training samples remains a prudent approach. The direction is to compare the effects of the quantity of training tasks on meta-training for detection performance. The authors' contributions include: Y.-K.L. conceptualization; Y.-K.L. methodology; Y.-K.L. software; Y.-K.L. validation; Y.-K.L. formal analysis; T.-Y.Y. investigation; Y.-K.L. original draft writing; Y.-K.L. review and editing writing; Y.-K.L. project administration; and Y.-K.L. funding acquisition. All authors have read and approved the manuscript as published.

Funding: The Ministry of Science and Technology in Taiwan funded this study under grant number MOST-109-2221-E-153-003.Data Availability Statement: Not relevant.

No conflicts of interest have been disclosed by the writers. The funders were not involved in the study's design, data collection, analysis, or interpretation, manuscript writing, or the choice to publish the findings.

6. References

[1] Yamasaki, T.; Shiohara, K. Self-Blended Images for Deepfake Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 19–20, 2022, New Orleans, Louisiana, USA, pp. 18720–18729.

[2] Thies, J.; Nießner, M.; Zollhöfer, M. Neural texture-based image synthesis is known as deferred neural rendering. 38, 1–12, ACM Trans. Graph. (TOG) 2019.

[3] Theobalt, C.; Nießner, M.; Stamminger, M.; Zollhofer, M.; Thies, J. Face2Face: Real-time facial recognition and rgb video reenactment. pp. 2387–2395 in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 26–30, 2016, Las Vegas, NV, USA.

[4] Thies, J. Face2Face: Reenacting faces in real time. 143–146 in IT-Inf. Technol. 2019, 61.

[5] Faceswap, 2018. retrieved on February 3, 2023, from https://github.com/MarekKowalski/FaceSwap.

[6] Deepfakes, 2018. Accessed on February 3, 2023, from https://github.com/deepfakes/faceswap.

[7] Riess, C.; Thies, J.; Nießner, M.; Verdoliva, L.; Rössler, A.; Cozzolino, D. Faceforensics++: Acquiring the ability to identify photos of altered faces. Proceedings of the IEEE/CVF International Conference on Computer Vision, 26 October–2 November 2019, Seoul, Republic of Korea, pp. 1–11.

[8] Bayar, B.; Stamm, M.C. Using a novel convolutional layer, a deep learning method for universal picture tampering detection. In Proceedings of the Fourth ACM Workshop on Multimedia Security and Information Hiding, Vigo, Spain, June 20–22, 2016; pp. 5–10.

[9] Verdoliva, L.; Poggi, G.; Cozzolino, D. An application to the identification of picture forgeries involves recasting residual-based local descriptors as convolutional neural networks. In Proceedings of the 5th ACM Workshop on Multimedia Security and Information Hiding, Philadelphia, PA, USA, June 20–22, 2017; pp. 159–164.

[10] Yamagishi, J.; Echizen, I.; Fang, F.; Nguyen, H.H. Multi-task learning for identifying and classifying movies and pictures with altered faces. arXiv:1906.06876, arXiv 2019.

[11] Zhou, Z.; Tajbakhsh, N.; Rahman Siddiquee, M.M.; Liang, J. Unet++: A nested u-net architecture for classifying medical images.

[12] Cheng, H.; Zhao, C.; Zhu, Y.; Cheng, C.; Feng, S.; Fan, Y.; Tang, Y. A Multi-Scale Adaptive Convolution Kernel Network and Multimodal Conditional Random Field-Based Change Detection Technique for Multi-Temporal Multispectral Images. Remote Sens. 14, 5368 (2022).