



KARADENİZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ SİBER GÜVENLİK LABORATUVARI



Multimedia Güvenliği

1. GİRİŞ

Deepfake terim olarak Deep learning ve Fake kelimelerinin birleşiminden oluşur ve yüz, mimik ve/veya sesi derin öğrenme yöntemleriyle gerçekçi biçimde üretmek ya da değiştirmek için kullanılan bir medya manipülasyon tekniğidir. Üretimde genellikle GAN ve otokodlayıcı tabanlı modeller kullanılır; bu sayede insan gözü için ikna edici içerikler üretmek mümkün hâle gelmiştir. Aynı zamanda açık kaynak yazılımlar ve mobil uygulamalar, teknik bilgi eşliğini dramatik biçimde düşürmüştür; neredeyse herkesin tek tıkla görsel/video üzerinde farklı türde dönüşümler yapabilmesini sağlamıştır.

Deepfake teknolojisi; filmlerde dublaj yapılması, vefat etmiş kişilerin filmlerde canlandırılması, tarihî karakterlerin görselleştirilmesi, kozmetik endüstrisinde uygulama sonuçlarının gösterilmesi ve çevrim içi toplantılarda bazı yararlı kullanımlar gibi pek çok faydalı uygulamayı da mümkün kılar.

Bununla birlikte tüm bu faydalı uygulamalara rağmen, deepfake video ve görüntüler; politik manipülasyon ve kamuoyu yönlendirmesi gibi ciddi zararlar ile kişisel mahremiyete yönelik tehditler oluşturabilecek önemli bir risk hâline gelmiştir. 2017’de “deepfakes” adlı bir Reddit kullanıcısı, bazı ünlülerin yüzlerinin pornografik videolardaki kişilerle değiştirildiği videolar paylaştı; bu içerik büyük yankı uyandırdı ve akabinde “deepfake” terimi özellikle videolarla özdeşleşti. 2019’da “DeepNude” adlı bir akıllı telefon uygulaması insanların fotoğraflarını çıplak hâle getiriyordu. Kısa süre sonra iOS için ZAO ve Android için FaceApp yayımlandı; bu uygulamalarla kullanıcılar tek tıkla bir görüntüde yüz değiştirme, saç/sakal ekleyip çıkarma, cinsiyet ve yaş değiştirme gibi çeşitli düzenlemeler yapabildi. 2021’de ise aktör Tom Cruise’a ait sahte videolar dolaşıma girerek kamuoyunda endişe yarattı.

Açık kaynak yazılımların ve çeşitli mobil uygulamaların varlığıyla, sahte görüntü ve videoların üretimi teknik bilgiye sahip olunmasa dahi neredeyse herkesin kolaylıkla yapabileceği bir noktaya gelmiştir. Üretilen sahte içerikler çoğu insan için son derece gerçekçi görüldüğünden, sahte olduklarının fark edilmesi güçleşmektedir.

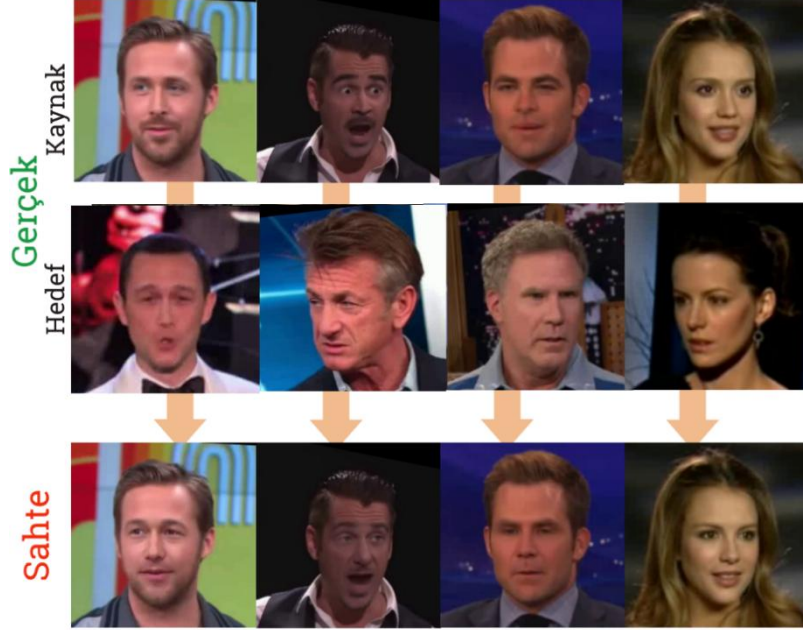
Deepfake içeriklerin üretim türleri; tüm yüz sentezi, kimlik değiştirme, özellik manipülasyonu, yüz geçişi ve ifade değişimi olmak üzere 5 başlıkta toplanmaktadır.

- A. Tüm yüz sentezi, var olmayan bir yüzün tamamen sıfırdan üretilmesi sürecidir. Genellikle ÇÜA tabanlı Style-GAN, Style-GAN2 ve StyleGAN2-Ada gibi modeller

kullanılmaktadır. ÇÜA'lar, birbirine karşı çalışan iki yapay sinir ağı olan üretici ve ayırt edici ağlardan oluşur. Üretici ağ, rastgele gürültüden gerçekçi görünen yüz görüntüleri üretmeye çalışırken, ayırt edici ağ bu görüntülerin sahte mi yoksa gerçek mi olduğunu ayırt etmeye çalışır. Bu karşılıklı öğrenme süreci, çok gerçekçi yüz görüntüleri oluşturulmasını sağlar. Üretilen görüntüler oldukça gerçekçidir ve insan gözüyle sahte olduğu anlaşılamayacak kalitededir.



- B. Kimlik değiştirme, bir videodaki kişinin yüzünün başka bir kişinin yüzüyle değiştirilmesidir. Bu teknik, son zamanlarda büyük toplumsal endişelere neden olan örnekleri nedeniyle en popüler yüz manipülasyon alanlarından biridir.



- C. Yüz geçişi, iki veya daha fazla kişinin yüz görüntüsündeki biyometrik özellikler kullanılarak yeni bir yapay biyometrik yüz oluşturulması işlemidir. Bu yöntem, kaynak yüzlerden alınan özelliklerin bir kombinasyonunu içerdiğinden, yüz tanıma sistemleri için ciddi bir tehdit oluşturmaktadır. Kimlik değişiminden farklı olarak yüz geçişinde, video seviyesinde sahte içerik oluşturmak yerine, görüntü düzeyinde sahte örnekler oluşturulur. Bu yöntemde üretilen sahte görüntüler, yüz tanıma sistemlerinin yanıltılması için kullanılabilir ve bu da güvenlik sistemlerinde büyük zorluklar yaratabilir.



- D. Özellik manipölasyonu, saç veya cilt rengi, cinsiyet, yaş, gözlük kullanımı gibi özelliklerin değiştirilmesi tekniğidir. ÇÜA tabanlı teknikler kullanılarak üretilen bu manipölasyonlar, piyasada son kullanıcının kullanımına sunulan birçok ücretli ve ücretsiz uygulama tarafından desteklenmektedir. Örneğin, FaceApp gibi mobil uygulamalar, kullanıcıların yüzlerinde çeşitli değişiklikler yapmalarına olanak tanır. Bu tür uygulamalar, kozmetik sektöründe ürünleri kullanmadan elde edilecek sonuçları görme gibi faydalı kullanımlar sağlayabilirken, aynı zamanda kötü niyetli kullanımlara da zemin hazırlayabilir.



- E. İfade değişimi, bir video veya görüntüdeki kişinin yüz ifadesinin değiştirilmesi işlemidir. Bu teknikte, kaynak videodaki kişinin yüz ifadeleri, hedef videodaki kişinin ifadelerine yansıtılır.



Deepfake videoların tespitinde kullanılan yöntemler aşağıda gösterilmektedir.



Kamera Parmak İzi (PRNU) Tabanlı Yöntemler: Her kamera sensörünün üretimden kaynaklı benzersiz bir “foto-tepki düzensizliği” (PRNU) izi vardır. Deepfake’te yüz bölgesi sonradan yerleştirildiğinde bu iz, videonun geri kalanıyla veya bilinen cihaz şablonuyla tutarsız hale gelir. Karelerden gürültü artık haritası (denoising sonrası residual) çıkarılır, yüz ROI’lerinde normalize korelasyon ile cihaz şablonu/arka plan PRNU’su karşılaştırılır. Tutarsızlık yüksekse “sahte” olasılığı artar. (Sıkıştırma/filtreleme/PRNU silme teknikleri bu tekniği zayıflatılabilir.)

Görsel Olgulara Dayalı Yöntemler: Deepfake yüzü ile özgün arka plan arasındaki birleşim hataları (dikiş, ten tonu/aydınlatma farkı, çözünürlük uyumsuzluğu) ve geometrik tutarsızlıklar (baş pozu–landmark uyumsuzluğu) sahteciliği ele verir. Yama tabanlı CNN’ler (VGG/ResNet) kompozit artefaktları öğrenir; Face X-ray benzeri temsiller, yüz-arka plan birleştirme sınırını doğrudan tahmin eder; 2B landmark’tan 3B baş pozu kestirip görüntüyle uyum test edilir. (Sınırın gizlendiği veya görüntü kalitesinin düşük olduğu durumlarda kırılabilir.)

Zamansal Tutarlılık Tabanlı Yöntemler: Video üretiminde kareler arası süreklilik bozulabilir; mikromimik, optik akış ve hareket örüntülerindeki zamansal kopukluklar tespit için kullanılır. Kare özellikleri CNN→RNN/LSTM zinciriyle veya 3D-CNN ile modellenir; yeni hatlarda ayrıştırılmış mekânsal-zamansal dikkat (Transformer) ve self-subtract gibi fark vurgulama mekanizmaları kareler arası uyumsuzlukları öne çıkarır. (Ağır sıkıştırma/düşük kalite genellemeyi zorlayabilir.)

Biyolojik Sinyallere Dayalı Yöntemler: Sahte yüzlerde doğal biyofizyolojik ritimler (göz kırpma düzeni, ciltteki rPPG nabız izi, dudak-ses eşzamanı) eksik veya yapay olabilir; bu da güçlü bir ayırt edici ipucudur. EAR ile göz kırpma zaman serisi, yüz bölgelerinden rPPG haritaları çıkarımı gibi yöntemler kullanılmaktadır. (Video kalitesi ve üreticilerin bu sinyalleri taklit etme yeteneği sınırlayıcıdır.)

Genel Ağ Tabanlı Yöntemler: Yüz görüntülerinden doğrudan “gerçek/sahte” sınıflandırması yapan CNN/Transformer/kapsül ağı gibi derin öğrenme modelleri, otomatik özellik öğrenerek yüksek doğruluk verebilir; ancak alan kayması ve aşırı uyum riski taşıyabilir.

2. DENEYE HAZIRLIK

Deney sırasında zaman kaybetmemek için aşağıdaki adımları laboratuvar öncesinde tamamlayınız.

Deneyde tüm kodlar anaconda ortamında çalıştırılacaktır. Bu nedenle anaconda kullanımı hakkında bilgi sahibi olunması gerekmektedir. Anaconda; Python dağıtımı + paket/ortam yöneticisi (conda) içeren bir platformdur. Farklı projeler için izole ortamlar oluşturarak kütüphane sürüm çatışmalarını önler. <https://www.anaconda.com/download/success> adresinden işletim sisteminize uygun sürümünü indirebilirsiniz. Laboratuvar kapsamında bilgisayarlarda anaconda önceden kurulu olacaktır. Anaconda kurulumunu gerçekleştirdikten sonra windows'ta başlata tıklayarak "Anaconda Prompt" yazarak anacondanın komut ekranı uygulamasını bulabilir ve başlatabilirsiniz. Anaconda içerisindeki işlemler bu ekrandan gerçekleştirilir. Anaconda'da deneyim kazanmak adına önceden kendi bilgisayarınızda kurup işlemler yapmanız önerilir.

3. DENEYİN YAPILIŞI

a) Anaconda'da ortam (environment) oluşturma

Neden? Her projeyi izole bir ortamda tutarak sürüm çakışmalarını önleriz.

Anaconda prompt ekranında aşağıdaki komutları yazarak yeni bir ortam oluşturulur. ortamadi kısmına kendi ortamınıza vermek istediğiniz ismi yazabilirsiniz.

```
~ conda create -n ortamadi python=3.10 -y
```

Ortam oluşturulduktan sonra aşağıdaki komutla ortam aktif edilir.

```
~ conda activate ortamadi
```

b) Gerekli paketleri kurma

Çalıştıracığımız kodlar için gerekli olan kütüphaneleri ortam içine kurmak için aşağıdaki komutları kullanmamız gerekmektedir.

```
~ pip install "torch==1.12.1" "torchvision==0.13.1" opencv-python==4.10.0.84  
mediapipe==0.10.14 numpy==1.26.4 pandas scikit-learn matplotlib tqdm
```

Deney süresince çalıştırılacak kodlar ve diğer dosyalar aşağıdaki adresten indirilecektir.

https://drive.google.com/file/d/1PH8BaL4ykwez1pRPsXXs_YDnwkXUU1ow/view?usp=sharing

c) Videolardaki kareleri (frame) çıkarma

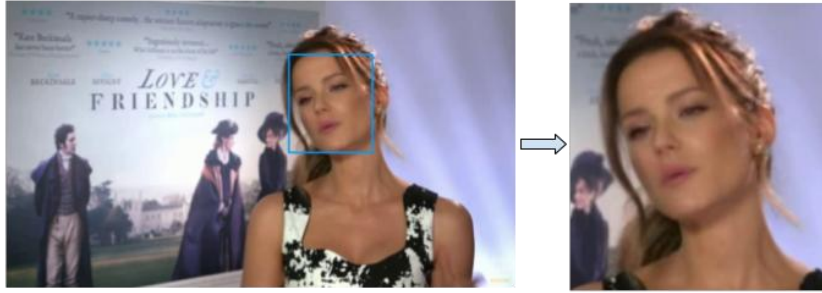
Videolar ardışık görüntü karelerinden oluşmaktadır. 30fps bir video saniyede 30 görüntü karesinden oluşan bir akışı temsil eder. 10sn uzunluğundaki 30fps bir video 300 görüntü karesinden oluşur. Derin öğrenme modelinde sınıflandırma testini yapabilmek için öncelikle

videodaki görüntü karelerini çıkarmamız gerekir. Sonrasında görüntüler için test işlemi yapılabilir.

```
~ python scripts/extract_frames.py --input_videos input_videos --output_dir frames --fps 1.0
```

Görüntü kareleri çıkarıldıktan sonra aşağıdaki örnekte solda gösterildiği gibi asıl odaklanılması gereken yüzün dışında çok fazla detay yer alabilir. Bu nedenle yüzün tespit edilerek sadece yüzün olduğu kısmın çıkarılması tespit işleminde başarıyı artıracaktır.

```
~ python scripts/crop_faces.py --input_dir frames --output_dir faces --margin 0.25
```



d) Frameleri test etme (model ile çıkarm)

Kırpılmış yüz karelerini modele verip video başına tek skor/etiket üreteceğiz. Anaconda prompt ekranındayken ilgili dizininde olduğumuzdan emin olduktan sonra `predict_videos.py` dosyasını çalıştırarak test işlemini gerçekleştireceğiz. Bunun için anaconda prompt ekranında aşağıdaki komutu çalıştıracacağız.

```
~ python scripts/predict_videos.py --faces_dir faces --weights checkpoints/capsule.pt --batch-size 32 --gpu -1
```

Kod çalışmasını tamamladığında `outputs` isimli klasör içerisine `results.csv` adında dosya oluşturacak ve dosyada her video için gerçek/sahte tahmini için belirlenen sonuç yazılacaktır.

e) Video için test sonucunu belirleme

Her video için üretilen test sonuçlarına bakarak sonuçları değerlendiriniz.

4. RAPOR SORULARI

- Deepfake içeriklerin faydalı kullanımları ve zararları hakkında yorumlarınızı örneklerle birlikte açıklayınız.
- Deepfake tespitinde var olan yöntemler içinde başarısının en iyi olacağını düşündüğünüz yöntem hangisidir? Nedenini açıklayınız.
- Deneyde yapılan işlemleri adım adım açıklayınız. Videolardan farklı sayılarda frame çıkararak sınıflandırma sonuçlarında değişiklik olup olmadığını belirleyin. Sonuçlara etkisini yorumlayınız.
- Hatalı sınıflandırılan videolar için neden hatalı sınıflandırma yapılmış olabileceği ile ilgilini yorumunuzu açıklayarak belirtiniz.