



TOBB Ekonomi ve Teknoloji Üniversitesi

Elektrik – Elektronik Mühendisliği Bölümü

ELE 567 – Haberleşme İçin Makine Öğrenmesi Projesi

Video Aktarımında Superresolution Kullanımı

Enes Sancak

201201004

İçindekiler

Giriş ve Proje Tanımı	3
Grace Modeli ve Benzerleri İncelemesi	3
GRACE: Sinirsel Codec ile Kayıp Toleransının Yeni Yüzü	4
DeepRS: FEC'e Derin Öğrenme ile Adaptasyon	4
DeepWiVe: JSCC ile End-to-End Video Gönderimi.....	5
Diğer Önemli Derin Öğrenme Tabanlı Yaklaşımlar	6
Proje Kapsamında Yapılan Çalışmalar	7
SR1 Modeli.....	7
SR2 Modeli.....	9
SR3 Modeli.....	11
SR4 Modeli.....	13
Sonuç	15
Kaynakça	16

Giriş ve Proje Tanımı

Günümüzde video tabanlı uygulamalar; bulut oyunları, uzaktan eğitim, sanal ve artırılmış gerçeklik gibi gerçek zamanlı sistemlerin ayrılmaz bir parçası haline gelmiştir. Bu uygulamaların performansı, büyük oranda video içeriğinin yüksek kaliteyle, düşük gecikmeyle ve mümkünse adaptif biçimde iletilmesine bağlıdır. Ancak ağ kaynaklarının sınırlılığı, özellikle kablosuz ortamlarda yaşanan dalgalanmalar, paket kaybı ve gecikme problemleri; geleneksel video kodlama tekniklerinin (örneğin H.264/H.265) sınırlarını zorlamaktadır. Bu bağlamda, son yıllarda derin öğrenme tabanlı video kodlama ve aktarım yöntemleri büyük ilgi görmektedir. Bu projede ise, gerçek zamanlı video iletimine yönelik geliştirilen GRACE framework'üne, video çözünürlük iyileştirme (super-resolution) görevini üstlenen sinirsel ağ modellerinin entegre edilmesi hedeflenmiştir.

Proje kapsamında 4 farklı SR (Super-Resolution) modeli sıfırdan tasarlanmış ve eğitimleri gerçekleştirmiştir. Modellerde zamanla mimari karmaşıklık, dikkat mekanizmaları ve kayıp fonksiyonları çeşitlendirilmiş, ardından performansı en dengeli model olan SR3, GRACE framework'üle entegre edilmiştir. Entegrasyon ve test süreçlerinin tamamı Python + PyTorch tabanlı Colab ortamında yürütülmüştür. Kodların tamamına ve model mimarilerine ait detaylara aşağıdaki bağlantı üzerinden erişim sağlanabilir:

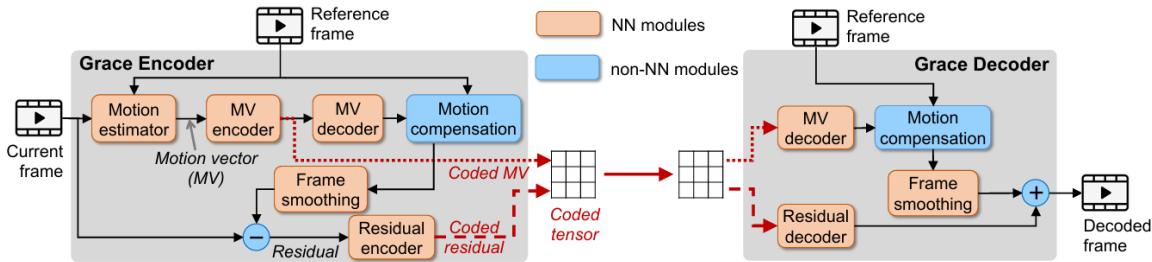
 [GraceWitSR \(GitHub\)](#)

Grace Modeli ve Benzerleri İncelemesi

Gerçek zamanlı video iletimi, çevrim içi toplantılar, bulut tabanlı oyunlar, uzaktan eğitim, sanal/gerçek artırılmış gerçeklik uygulamaları gibi sayısız alanda kullanıcı deneyimini doğrudan etkileyen bir teknolojidir. Bu tür uygulamalar, ağ üzerinden yüksek kaliteli video verisinin düşük gecikmeyle ve kesintisiz biçimde iletilmesini gerektirir. Ancak, kablosuz ağlardaki dalgalı bant genişliği, yüksek gecikme varyansı ve paket kaybı oranları, geleneksel video kodlama çözümlerini zorlayıcı hale getirmektedir. Özellikle sıkıştırılmış video paketlerinin zamanında ulaşamaması, karelerin bozulmasına veya tamamen atlanmasına yol açar. Bu nedenle, paket kaybına dayanıklı, ağ koşullarına uyum sağlayabilen ve gerçek zamanlı çalışabilen çözümlere olan ihtiyaç giderek artmaktadır.

Bu bağlamda, son yıllarda derin öğrenme temelli video kodlama ve iletim yaklaşımları ön plana çıkmıştır. Bu yöntemler, geleneksel codec'lerin yerine geçerek veya onların işlevselliğini artırarak paket kaybı toleransı, sıkıştırma verimliliği ve görsel kalite alanlarında önemli gelişmeler sunmaktadır. Literatürde bu alanda öne çıkan üç temel yaklaşım kategorisi bulunmaktadır: (1) sinirsel video codec sistemleri, (2) derin öğrenme destekli ileri düzey FEC (Forward Error Correction) stratejileri ve (3) ortak kaynak-kanal kodlama (Joint Source-Channel Coding, JSCC) yöntemleri başlıca gelen yöntemlerdir.

GRACE: Sinirsız Codec ile Kayıp Toleransının Yeni Yüzü



Şekil 1: Grace Enkoder ve Dekoder Modeli

GRACE (Generative Resilient and Adaptive Codec for Encoding) framework'ü [Cheng et al., 2024], derin öğrenme tabanlı bir video codec sistemidir ve kayıplı ağlar üzerinde yüksek kaliteli video iletimi sağlamayı hedefler. GRACE'in en belirgin özelliği, encoder ve decoder bileşenlerinin birlikte, çeşitli kayıp senaryoları altında ortaklaşa eğitilmesidir. Bu yaklaşım, klasik FEC ve error concealment tekniklerinden temel olarak farklıdır. Encoder, bilgi temsillerini kayıplara karşı daha dayanıklı olacak şekilde optimize ederken, decoder tarafı da eksik paketlerle anlamlı kareler üretmeye yönelik olarak eğitilmektedir. Bu çift taraflı öğrenme, sinirsız codec mimarilerinin geleneksel yapılarına göre çok daha yüksek başarı oranına ulaşmasına olanak tanır.

GRACE'in eğitim aşamasında, tensör seviyesinde rastgele sıfırlama (random zeroing) yöntemiyle paket kaybı senaryoları simüle edilir. Encoder tarafından üretilen tensör çıktılar, rasgele böülümlere ayrılarak paketleştirilir. Decoder ise eksik gelen böülümleri sıfırla doldurarak yeniden yapılandırma işlemi yapar. GRACE, geleneksel Reed-Solomon temelli FEC veya HEVC'deki tile/slice partitioning gibi çözümlerden farklı olarak kayıp karşısında sabit oranda değil, sürekli degrade olan bir kalite eğrisi sunar – bu "graceful degradation" özelliği literatürde önemli bir avantaj olarak kabul edilir.

Ek olarak, GRACE içerisinde yer alan "dynamic resynchronization" protokolü sayesinde encoder ve decoder senkronizasyonu, paket kaybı durumunda minimum veri yüküyle sağlanır. Encoder, decoder'dan gelen feedback sinyaline göre referans kare hafızasını güncelleyebilir. Bu, hem ek yeniden iletim ihtiyacını ortadan kaldırır hem de senkronizasyon sorunlarını çözerek kesintisiz video sunumuna olanak tanır.

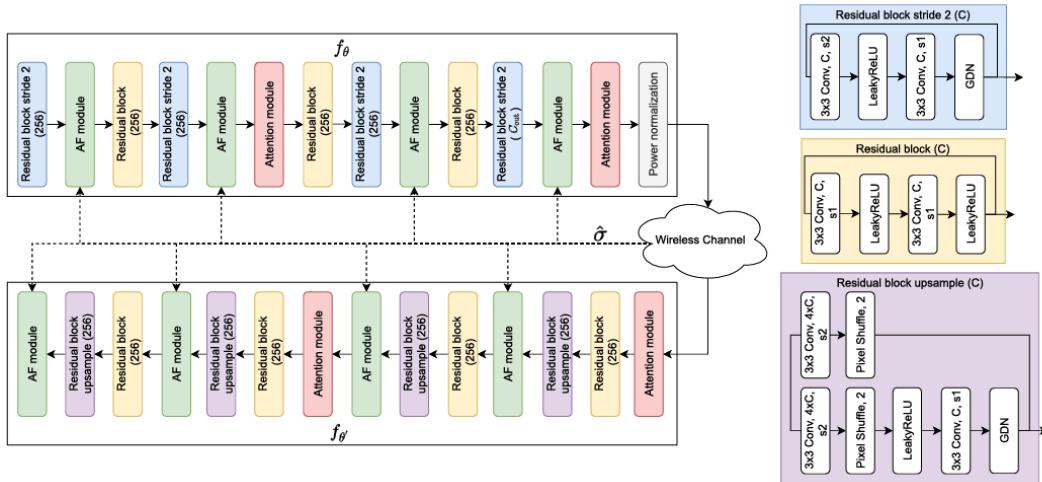
DeepRS: FEC'e Derin Öğrenme ile Adaptasyon

Cheng et al. tarafından geliştirilen DeepRS (Deep-learning-based Network-Adaptive FEC for Real-Time Video Communications) [Cheng et al., 2020], klasik Reed-Solomon FEC sistemlerini modernize etmeye yönelik bir çözümüdür. Temel fikir, alıcıdan gelen gecikmeli geri bildirimleri kullanarak, gelecekteki paket kaybı olasılıklarını bir LSTM (Long Short-Term Memory) ağı ile tahmin etmektedir. Böylece encoder tarafında uygulanacak fazlalık oranı dinamik biçimde güncellenir.

DeepRS'in fark yarattığı nokta, fazlalık oranlarının statik değil tahmine dayalı olarak ayarlanmasıdır. Deneyel sonuçlar, DeepRS'in klasik sabit FEC sistemlerine kıyasla sabit bant genişliğinde %70'e kadar daha fazla paket kurtarımı sağladığını göstermektedir. Ancak dikkat çeken sınırlılık, decoder tarafının öğrenme sürecine dahil edilmemiş olmasıdır. Bu durum, GRACE gibi çift taraflı optimize edilmiş sistemlere kıyasla bütüncül kalite artışını sınırlıtmaktadır.

DeepRS benzeri yapılar, özellikle VoIP ve düşük bitrate'lı canlı yayın uygulamalarında ağ fazlalığını minimal tutarak kaliteli iletim için uygulanabilir çözümler sunar. Ancak, görüntü kalitesi açısından kayıpların olduğu senaryolarda (örneğin 30%+ kayıplı UDP) GRACE benzeri encoder-decoder uyumlu yapılar daha başarılıdır.

DeepWiVe: JSCC ile End-to-End Video Gönderimi



Şekil 2: Deep WiVe

DeepWiVe (Deep-Learning-Aided Wireless Video Transmission) [Tung & Gündüz, 2022], son yıllarda popülerlik kazanan JSCC (Joint Source-Channel Coding) yaklaşımının video gönderimi özelindeki uygulamalarından biridir. Bu sistemde, video verisi doğrudan kanal girişine dönüştürülür; klasik anlamda sıkıştırma (source coding) ve kanal kodlaması (channel coding) aşamaları ayrı ayrı değil, bir arada gerçekleştirilir. DeepWiVe, encoder ve decoder yapılarında derin sinir ağları kullanır ve kanal üzerinden analog olarak öğrenilmiş temsiller gönderilir.

Sistemde ayrıca kare başına değişken bant genişliği tahsisini öğrenen bir RL (Reinforcement Learning) tabanlı kontrol katmanı bulunmaktadır. Bu katman, kanal kalitesine göre her bir kare için ayrılacak kaynak miktarını optimize eder. DeepWiVe, hem H.264 + LDPC hem de H.265 + LDPC sistemlerine kıyasla MS-SSIM açısından ortalama %6'ya kadar daha iyi sonuçlar sunmaktadır.

Ancak, JSCC sistemlerinin yaygınlaşmasının önündeki en büyük engel açıklanabilirlik (explainability) ve sistem karmaşıklığıdır. Özellikle kablosuz kanal modelleri ile birlikte eğitilen ağların eğitimi, çok sayıda hiperparametre ve kanal simülasyonu gerektirdiğinden mühendislik açıdan karmaşıklıdır. Ayrıca, dijital yayın altyapılarında analog bilgi akışının sistem entegrasyonu, yasal düzenlemeler ve codec standardizasyonları nedeniyle sınırlı kalabilir.

Düzenleme 2: Diğer Önemli Derin Öğrenme Tabanlı Yaklaşımlar

Bu alanda GRACE, DeepRS ve DeepWiVe dışında literatürde yer alan başka yenilikçi sistemler de bulunmaktadır:

Salsify (Yan et al., 2018), ağ koşullarına göre gerçek zamanlı olarak kodlama ve iletim parametrelerini güncelleyebilen bir sistemdir. Bu çalışma, video bit akışlarını ağın anlık RTT ve kayıp oranlarına göre uyarlamak için klasik codec'leri kullanmak yerine uçtan uca ayarlanabilir bir mimari önerir. Özellikle düşük gecikme kritik uygulamalarda, video karelerinin paketlenme biçimini de dinamik olarak şekillendirir. Salsify, "receiver-driven" bir yaklaşım benimseyerek alıcı tarafın oynatma tampon durumuna göre kodlama hızını düzenler.

NeRV (Chen et al., 2021), geleneksel kare dizileri yerine her kareyi ayrı bir sinir ağı çıktısı olarak temsil eden yeni bir yaklaşım getirir. Bu mimaride, bir sinir ağı video verisinin tamamını sıkıştırılmış bir formatta taşıır. Zaman içinde her kare, giriş zamanı indeksine göre yeniden oluşturulur. NeRV, düşük bitrate senaryolarında bile sabit kalite sunarken, aynı zamanda hafıza kullanımını da minimize eder. Bu yönyle hem saklama hem de iletim optimizasyonu açısından yenilikcidir.

FVC (Lu et al., 2019), alanında önemli bir çalışmındır. Karelere arası zaman korelasyonunu kullanarak, her kare için motion + residual bilgi üretir. Bu bilgi encoder-deep CNN yapısı üzerinden sıkıştırılır. Motion kompanzasyonu, geleneksel optik akış yöntemlerine kıyasla doğrudan sinir ağı ile tahmin edilmekte ve residual bileşenleri ile birlikte birlikte kodlanmaktadır.

Scale-Space Flow (SSF) ve DVC (Wu et al., 2018) gibi çalışmalar, derin öğrenme tabanlı video sıkıştırma alanında optik akış tahmini ve temporally residual encoding'i birleştirerek yüksek verimlilikte sonuçlar elde etmektedir. Özellikle DVC sistemi, klasik B-frame yapısına benzer şekilde ileri ve geri akışlarla kare tahmin eder ve residual ağları ile kaliteyi yükseltir.

Bu çalışmaların tamamı, sinir ağlarının hem kodlama hem kanal ortamı hem de yeniden yapılandırma sürecinde kullanılmasının ne denli etkili olduğunu ortaya koymaktadır. Her biri farklı bir tasarım tercihi getirirken, bu tercihlerin başarı oranı uygulamaya ve sistem gereksinimlerine göre değişmektedir. Ancak ortak nokta, klasik video iletim mimarilerinin ötesine geçerek daha esnek, uyarlanabilir ve kalite odaklı çözümler sunmalarıdır.

Proje Kapsamında Yapılan Çalışmalar

Proje kapsamında kullanılan veri seti, Microsoft Research tarafından geliştirilen ve video anlama ile açıklama gibi multimodal görevler için yaygın olarak kullanılan MSR-VTT (Microsoft Research Video to Text) veri kümeleridir. Toplamda 10.000 adet kısa web videosu içeren bu küme, 20 farklı temayı kapsayan zengin içerik çeşitliliğiyle derin öğrenme modellerinin eğitimi için güçlü bir temel sunmaktadır. Her biri yaklaşık 15 saniye uzunlığında ve 240p çözünürlükte sunulan videolar, çeşitli popüler arama motorlarından toplanmış olup; bu özellikleri sayesinde özellikle video süper çözünürlük (super-resolution) gibi görsel kalite iyileştirme çalışmalarında düşük maliyetli ve kısa süreli deneysel prototiplemeler (proof-of-concept) için oldukça uygundur. Bu projede, yalnızca video bileşenleri kullanılmış; doğal dil açıklamaları dahil edilmemiştir. Ayrıca model eğitiminin hızlandırılması ve donanım maliyetinin azaltılması amacıyla, her videonun yalnızca ilk 2 saniyesine ait kareler seçilerek özel bir alt veri seti oluşturulmuş ve tüm modeller için yaklaşık %70 eğitim, %10 doğrulama ve %20 test oranlarında bir bölme uygulanmıştır.

Proje kapsamında basitten karmaşığa doğru 4 ayrı SR modeli eğitilmiştir. Bu modeller eğitirken başlangıçta yapılan hatalar göz önüne alınıp modelde iyileştirmeler yapılmaya çalışılmıştır. Projenin kapsamında en ideal modele ulaşılmasada en azından ortanın üstünde 240p videolar için çalışan bir model (SR3 modeli) eğitilebilmiştir. Sırasıyla modelleri inceleyip yaşanan teknik zorluklar ve bir sonrakinde bunu düzeltmek amacıyla yapıaln geliştirmeler ve bazı sonuçlar verilmiştir. Modellerin tamamı Colab ortamında Cuda platformu kullanılarak eğitilmiştir.

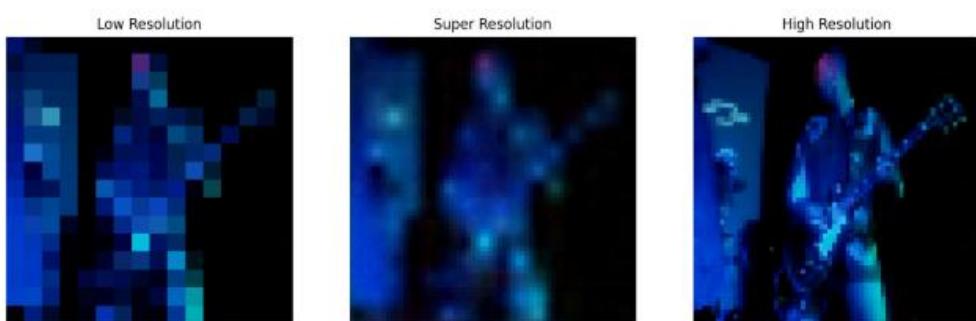
SR1 Modeli

SR1 modeli, video süper çözünürlük (VSR) alanında temel seviye bir referans model olarak tasarlanmıştır. Amaç, düşük çözünürlüklü video karelerini daha yüksek çözünürlüklü ve görsel olarak daha kaliteli karelere dönüştürmektir. Bu bağlamda SR1, hafif bir yapıya sahip olup gerçek zamanlı çalışmaya elverişli olmasına dikkat çeker. Model, temel olarak residual öğrenme yaklaşımını benimseyen konvolüsyonel katmanlar ve piksel tabanlı yeniden örneklemeye (pixel shuffle) mekanizması üzerine kuruludur. Yapısında dört adet hızlı residual blok ve tek adımda 4x upscaling yapan bir pixel shuffle modülü bulunur. Bu sayede hem bilgi aktarımı etkin bir şekilde sağlanmakta hem de ağ derinliği kontrol altında tutularak eğitim süreci basitleştirilmektedir.

Katman	Tip	Parametreler
Input	-	[batch, frames, 3, H/4, W/4]
Conv1	Conv2d	in_channels=3, out_channels=32, kernel_size=3, padding=1
ResBlock1	FastResidualBlock	num_channels=32
ResBlock2	FastResidualBlock	num_channels=32
ResBlock3	FastResidualBlock	num_channels=32
ResBlock4	FastResidualBlock	num_channels=32
Upsample	Conv2d + PixelShuffle	in_channels=32, out_channels=32*(4^2), kernel_size=3, padding=1
Conv2	Conv2d	in_channels=32, out_channels=3, kernel_size=3, padding=1

Şekil 4: SR1 Modeli Yapısı

Modelin toplam parametre sayısı yalnızca ~223K seviyesindedir. Bu kompakt yapı, SR1'i mobil ve gömülü sistemlerde kullanım açısından cazip hale getirir. Eğitim sürecinde kullanılan veri seti, MSR-VTT veri kümesinin 7000 video içeren alt kümesidir. Her videodan yalnızca ilk 2 saniyedeki kareler alınarak düşük çözünürlükte (64×64) giriş, yüksek çözünürlükte (256×256) hedef çiftleri oluşturulmuştur. Eğitim sırasında yalnızca MSE (Mean Squared Error) kayıp fonksiyonu kullanılmış, optimizasyon olarak Adam tercih edilmiştir. Ancak bu yaklaşım, modelin özellikle metin gibi yüksek frekanslı detayları ve ince kenarları yeniden üretmede ciddi kısıtlamalara sahip olduğunu ortaya koymuştur. Deneysel sonuçlar modelin özellikle yazıların bulunduğu bölgelerde karakter okunabilirliğinde düşüş, kenar keskinliklerinde kayıplar ve dokularda bulanıklık gibi sınırlamalar gözlemlenmiştir. Bu durum, yalnızca piksel farkına dayalı bir kayıp fonksiyonunun süper çözünürlük gibi karmaşık bir görev için yeterli olmadığını ve daha sofistike, insan algısına daha yakın kayıp fonksiyonlarının (perceptual loss, SSIM loss gibi) kullanılması gerektiğini ortaya koymuştur. Yani sıradan MSE fonksiyonuyla model adeta bir ortalama alma görevini öğrenmektedir. Bu modelin eğitimi bir kaç saat sürmüştür ama sonrasında kullanıcı eliyle iptal edilmiştir. Nedeni ise loss fonksiyonunun başlangıçtan itibaren oldukça düşük olmasındandır. İlk çalışmalardan birisi olduğu için train ve validasyon kayıplarının grafik verisi elde edilememiştir.



Şekil 3: SR1 Model Test Video Karesi Sonuç

SR2 Modeli

SR2 modeli, SR1'in sunduğu temel yapıyı koruyarak üzerine önemli iyileştirmeler eklenmiş ikinci nesil bir video süper çözünürlük (VSR) mimarisidir. SR1'in en temel eksikliklerinden biri olan yüksek frekanslı detayların kaybı ve yapay detayların üretilememesi gibi sorunları azaltmak amacıyla SR2'de hem mimari hem de öğrenme algoritması düzeyinde çeşitli yenilikler gerçekleştirilmiştir. Model, 8 adet residual blok, SE (Squeeze-and-Excitation) mekanizması, Batch Normalization ve Dropout gibi gelişmiş katmanlardan oluşur. Bu yapı, kanal bazlı dikkat mekanizması sayesinde önemli özelliklerin daha iyi öğrenilmesini sağlamış, dropout katmanları ise modelin genelleme kapasitesini artırarak overfitting riskini azaltmıştır.

Modelin çıkış boyutu 2x upsampling ile 240x240 çözünürlüğe ulaşmaktadır. Progressive upsampling ve pixel shuffle yapısı sayesinde keskin geçişler daha iyi yakalanmakta ve interpolasyona dayalı bulanıklık azaltılmaktadır. SR1'e kıyasla toplam parametre sayısı %50'den fazla azaltılarak yaklaşık 117K seviyesine düşürülmüş, buna rağmen kalite kaybı yaşanmamış, aksine kalite artışı sağlanmıştır. Bu durum, SR2'nin hem daha verimli hem de daha etkili bir çözüm sunduğunu göstermektedir.

Katman Tipi	Parametreler	Çıktı Boyutu
Input	-	(B, 3, H, W)
Initial Conv	3→24, k=3, s=1, p=1	(B, 24, H, W)
Residual Blocks (8x)	24→24, k=3, s=1, p=1	(B, 24, H, W)
Conv Mid	24→24, k=3, s=1, p=1	(B, 24, H, W)
Upscale Conv1	24→96, k=3, s=1, p=1	(B, 96, H, W)
PixelShuffle	scale=2	(B, 24, 2H, 2W)
Upscale Conv2	24→24, k=3, s=1, p=1	(B, 24, 2H, 2W)
Upscale Conv3	24→3, k=3, s=1, p=1	(B, 3, 2H, 2W)
AdaptiveAvgPool	(240,240)	(B, 3, 240, 240)

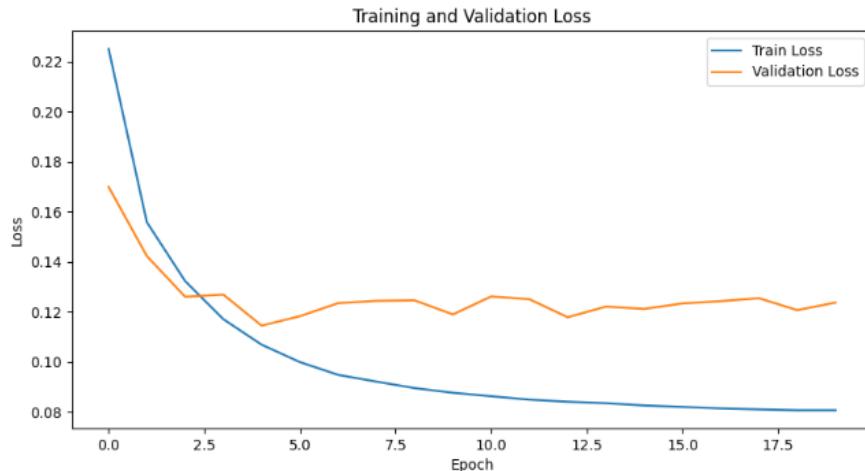
Şekil 5: SR2 Model Yapısı

SR2'nin en dikkat çekici yönlerinden biri, kayıp fonksiyonu (loss function) tasarımidır. MSE tabanlı tekil kayıp yaklaşımı yerine, L1 loss, Perceptual loss ve SSIM loss fonksiyonlarının ağırlıklı toplamından oluşan bileşik bir yapı benimsemistiştir. Bu üçlü kayıp yapısı sayesinde hem piksel seviyesinde hata azaltılmış, hem de görüntülerin yapısal ve semantik benzerliği korunmuştur. L1 kaybı temel doğruluğu sağlarken, perceptual kayıp VGG19'un ilk katmanları üzerinden görsel kaliteyi, SSIM ise yapısal benzerliği optimize etmiştir.

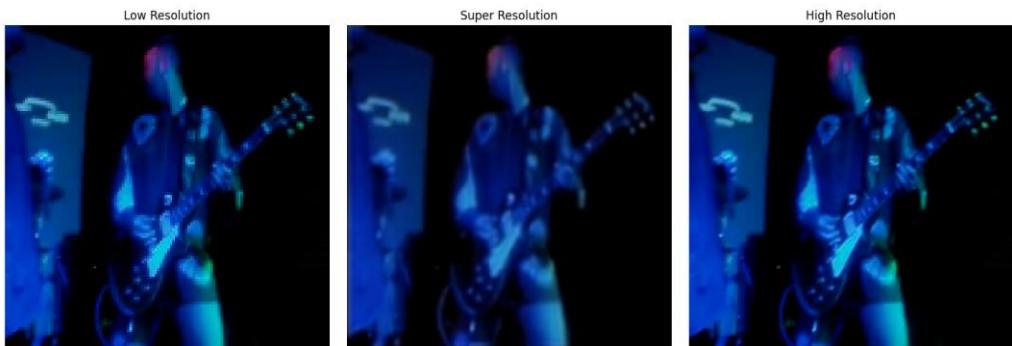
Eğitim süreci 100 epoch boyunca Adam optimizer ile yürütülmüş, ReduceLROnPlateau scheduler ile öğrenme oranı dinamik biçimde ayarlanmıştır. Validation set performansına göre early stopping mekanizması devreye alınmış, modelin aşırı öğrenme yapmadan en verimli noktada durdurulması sağlanmıştır. Eğitim boyunca hem training hem de validation loss'ta belirgin düşüşler gözlemlenmiş, özellikle SR1'e kıyasla PSNR ve SSIM skorlarında anlamlı iyileşme elde edilmiştir.

Sonuç olarak, SR2 modeli, SR1'in yapı taşlarını alarak üzerine daha güçlü öğrenme mekanizmaları, dikkat katmanları ve loss fonksiyonu entegrasyonları ekleyerek hem daha kompakt hem de daha başarılı bir mimari ortaya koymustur. Görsel kalite bakımından daha net kenar çizimleri ve yapısal bütünlüğü daha yüksek kareler üretmiş, özellikle insan gözüyle fark edilebilir bulanıklık ve yazı bozulması problemlerinde ciddi iyileştirmeler sunmuştur. Ancak ayrıntılı şekiller gibi video içi noktalarda halen daha o bulanıklılık etkisi devam etmekte olduğundan ve model eğitilirken gradyan

silinmesi (nan olması) sorunu yaşandığından bir sonraki modelde biraz daha farklı iyileştirmelere gidilmiştir. Aynı zamanda belirli bir epoch noktasından sonra validasyon data setindeki kayıp değeri sabit kalırken eğitim setindeki kayıp değeri azaldığı için overfittingi önlemek amacıyla daha önceden ayarlanmış bir patience noktasından sonra early stopping devreye girerek eğitimi bitirmiştir. Sonuçlar aşağıda görülebilir. Ayrıca bu modelden itibaren Grace'in giriş sınırı 128p olduğu için scale faktörü 1.85 seçilerek modeller eğitilmiştir.



Şekil 7: SR2 Eğitim Süreci Kayıp



Şekil 6: Test Video Karesi Sonuç (Genel)



Şekil 7: Test Video Karesi Sonuç (Yazılı ve Şekilli)

SR3 Modeli

SR3 modeli, video süper çözünürlük (VSR) problemine yönelik olarak SR2'nin üzerine inşa edilmiş, yapısal olarak daha derin ve dikkat mekanizmaları ile zenginleştirilmiş bir derin öğrenme mimarisidir. SR2 modelinin kazandırdığı başarıların üzerine çıkabilmek amacıyla SR3'te hem model kapasitesi hem de ağ derinliği önemli ölçüde artırılmıştır. Residual blok sayısı 8'den 10'a çıkarılmış, kanal sayısı 24'ten 48'e yükseltilmiş ve modelin toplam parametre sayısı yaklaşık 117K'dan 655K'ya çıkarılmıştır. Bu kapsamda SR3, SR2'ye kıyasla daha büyük ve daha karmaşık yapıda bir model olarak öne çıkmaktadır.

SR3 modelinde kullanılan temel mimari yapı, konvolüsyonel katmanlar, residual bloklar, SE (Squeeze-and-Excitation) kanal dikkat mekanizması ve yeni olarak eklenen Spatial Attention katmanlarıyla desteklenmiştir. Residual bloklardan sonra eklenen bu dikkat katmanları, özellikle videolardaki yazı ve küçük detayların algılanması için kritik katkılarda sunmuştur. Her residual blok içinde Batch Normalization, ReLU aktivasyonu ve Dropout (0.2 oranında) yer almaktadır; bu yapı hem düzenli gradyan akışı sağlamış hem de overfitting'i önlemeye yardımcı olmuştur.

Katman Tipi	Parametreler	Çıktı Boyutu
Input	-	(B, 3, H, W)
Initial Conv	$3 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Initial Conv2	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Residual Blocks (10x)	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Channel Attention (5x)	-	(B, 48, H, W)
Conv Mid	$48 \rightarrow 96 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Upscale Conv1	$48 \rightarrow 192, k=3, s=1, p=1$	(B, 192, H, W)
PixelShuffle	scale=2	(B, 48, 2H, 2W)
Upscale Conv2	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, 2H, 2W)
Upscale Conv3	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, 2H, 2W)
Final Conv	$48 \rightarrow 3, k=3, s=1, p=1$	(B, 3, 2H, 2W)

Şekil 8: SR3 Modeli Yapısı

Önemli bir diğer gelişme, kayıp fonksiyonunun SR2'de kullanılan kombinasyonun korunarak optimize edilmesidir. SR3'te yine L1 Loss ($\alpha=0.7$), Perceptual Loss ($\beta=0.2$) ve SSIM Loss ($\gamma=0.1$) ağırlıklarıyla birlikte kullanılmıştır. Görüntülerin yapısal ve semantik düzeyde daha kaliteli üretilmesini sağlayan bu bileşik kayıp yapısı sayesinde, SR2'ye göre PSNR (~33.2 dB) ve SSIM (~0.93) skorlarında artış elde edilmiştir. Eğitim sürecinde, klasik Adam yerine daha yeni ve hafif adaptasyon stratejileri içeren Lion optimizator tercih edilmiştir. Öğrenme oranı, her 10 epoch'ta bir yarıya düşen StepLR ile düzenlenmiştir, böylece uzun vadeli eğitim stabilitesi artırılmıştır.



Şekil 9: SR3 Eğitim ve Validasyon Kaybı

Düşük çözünürlüklü kareler 128x128 boyutuna çıkarılmış, hedef çıktılar ise 240x240 olarak belirlenmiştir. Bu geçiş, modelin ince detayları daha iyi yakalamasını sağlamış ancak bellekte daha fazla yük oluşturmuştur. CUDA bellek ayarları optimize edilmiş, frame caching ve autocast gibi tekniklerle eğitim süreci hızlandırılmış ve GPU kaynak kullanımı azaltılmıştır. Bununla birlikte, modelin parametre sayısının artması eğitim süresini yaklaşık 20 saatler civarına çıkarmıştır.

Sonuç olarak SR3 modeli, hem yapısal tasarımları hem de öğrenme mekanizmaları açısından SR2'ye kıyasla anlamlı bir sıçrama sunmuş, özellikle görsel detayların ve karmaşık içeriklerin çözünürlük iyileştirmesinde daha başarılı sonuçlar üretmiştir. Ancak hala dhaao kunabilir metinleri çıkartmakta zorlandığından SR4'te bu problem üstüne gidilmeye çalışılmıştır.



Şekil 10: SR3 Modeli Video Kare Test

SR4 Modeli

SR4 modeli, video süper çözünürlük problemini metin tanıma ve ince detaylar üzerine optimize edilmiş bir yapıyla ele almak amacıyla, SR3 mimarisine ek olarak Text Attention katmanlarıyla genişletilmiş son derece yoğun bir ağdır. SR3'te başarıyla uygulanan kanal (SE) ve uzamsal (Spatial) dikkat mekanizmalarının üzerine, metin içeriğini daha etkin öğrenmek amacıyla Text Attention yapıları eklenmiştir. Bu sayede yazı gibi yüksek frekanslı ve yapısal olarak belirgin öğeleri daha doğru biçimde yakalamak hedeflenmiştir. Ancak bu agresif detay odaklı mimari, bazı beklenmedik yan etkiler doğurmuştur.

Katman Tipi	Parametreler	Çıktı Boyutu
Input	-	(B, 3, H, W)
Initial Conv	$3 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Initial Conv2	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Residual Blocks (10x)	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Channel Attention (5x)	-	(B, 48, H, W)
Text Attention (10x)	-	(B, 48, H, W)
Spatial Attention (10x)	-	(B, 48, H, W)
Conv Mid	$48 \rightarrow 96 \rightarrow 48, k=3, s=1, p=1$	(B, 48, H, W)
Upscale Conv	$48 \rightarrow 192, k=3, s=1, p=1$	(B, 192, H, W)
PixelShuffle	scale=2	(B, 48, 2H, 2W)
Upscale Conv2	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, 2H, 2W)
Upscale Conv3	$48 \rightarrow 48, k=3, s=1, p=1$	(B, 48, 2H, 2W)
Final Conv	$48 \rightarrow 3, k=3, s=1, p=1$	(B, 3, 2H, 2W)



Şekil 11: SR4 Modeli Yapısı ve Eğitim

SR4'ün en önemli sorunu, çıktı görüntülerde görülen belirgin renk kaybıdır. Bu durum, modelin detay üretmede başarılı olmasına rağmen, genel görsel gerçekçilikte ciddi bozulmalara yol

açmaktadır. Sorunun kökeninde birkaç mimari faktör yatkınlıkta: Text Attention Mekanizması, renk kanallarına değil yapısal farklılığa odaklanmakta, böylece semantik olarak önemli ancak renk açısından tutarsız çıktılar üretmektedir. Aşırı Derinlik, modelin öğrenme sürecinde gradyan geçişini zorlaştırmakta ve öğrenilmesi gereken renk ilişkilerini maskeler hâle getirmektedir. Loss Ağırlıkları, özellikle Perceptual Loss'un SR3'e kıyasla azaltılmış olması ($0.2 \rightarrow 0.15$), L1 ağırlığının artırılması ($0.7 \rightarrow 0.8$), renk ve dokunun korunmasını zorlaştırmıştır. Batch Normalization'ın fazlalığı ve ReLU gibi aktivasyon fonksiyonlarının sınırlayıcı doğası da negatif değerleri sıfıra indirgerek renk tonlarının hassas işlenmesini engellemiştir. Elde edilen görsellerde bu sorun açıkça gözlemlenmiştir: Metin gibi yapılar netleşmiş ancak renk doygunluğu ciddi şekilde azalmış, gerçekçi renk geçişleri bozulmuştur. Özellikle SR4'ün bu haliyle GRACE gibi gerçek zamanlı ve kullanıcı odaklı video aktarım sistemlerinde kullanımı büyük ölçüde problemlü hale gelmiştir.



Şekil 12: SR4 Modeli Test

Sonuç

Bu çalışma kapsamında, detay odaklı mimari yapısına rağmen ciddi renk kaybı ve yapısal bozulmalar sergileyen SR4 modeli yerine, daha dengeleyici ve stabil bir çözüm sunan SR3 modeli, GRACE framework'ü ile entegrasyon için tercih edilmiştir. SR3, SR4 kadar agresif bir detay ayrıştırma kapasitesine sahip olmamakla birlikte; renk bütünlüğünü koruyan, görsel kalite açısından daha tutarlı sonuçlar üreten ve sistemin mobil/düşük gecikmeli doğasına daha uygun çalışan bir yapıya sahiptir. Ayrıca SR3'ün daha düşük parametre sayısı (~655K) ve kısa sürede eğitilebilir oluşu, GRACE'in gerçek zamanlı veri iletimine yönelik yapısı ile teknik olarak uyumlu bir entegrasyon sürecine imkân tanımıştır.

GRACE + SR3 entegrasyonu, özellikle görsel kaliteyi yükseltmek ve sıkıştırılmış video çıktısını kullanıcı tarafında yeniden yapılandırmak amacıyla kurgulanmıştır. Bu entegrasyon, belirli düzeyde gürültü azaltma, kenar netleştirme ve detay yeniden yapılandırma görevlerinde başarı sağlamış; düşük çözünürlükten elde edilen karelerin algısal kalitesinde gözle görülür bir iyileşme sunmuştur. Bununla birlikte, proje sınırlı zaman ve kaynak planlaması sebebiyle tam anlamıyla optimize edilememiştir. Özellikle, zaman eksenindeki tutarlılık, kareler arası geçişlerdeki gürültü kaynaklı bozulmalar ve çözünürlük ölçeklemesi sırasında oluşan zaman uyumsuzlukları çözümlemesi gereken temel sorunlar olarak gözlemlenmiştir.

Gelecek çalışmalarında hedef, bu tür bir SR mimarisinin yalnızca görsel kaliteyi artırmakla kalmayıp, aynı zamanda video aktarımında bant genişliğini azaltarak daha kesintisiz ve kaliteli hizmet sunacak şekilde evrilmesidir. Ancak bu amaç doğrultusunda mimarinin hesaplama karmaşıklığı ve model boyutu, gerçek zamanlı sistemler için donanım kısıtlarını aşmayacak seviyede optimize edilmelidir. Bu bağlamda, sadece SR3 değil, aynı zamanda GRACE framework'ünün de daha hafif, parametre açısından sadeleştirilmiş versiyonları önerilebilir. Böyle bir yaklaşım hem kaliteyi hem de verimliliği aynı anda hedefleyen hibrit çözümlerin önünü açacaktır.

Aşağıda, SR3 modeli ile GRACE framework'ü entegrasyonundan elde edilen bir test videosunun giriş ve çıkış kareleri sunulmuştur. Görsel çıktıların temel düzeyde başarılı olduğu gözlemlense de, güçlü donanımlar ve zaman-temelli bağımlılıkları öğrenen yeni nesil modellerle desteklendiğinde, gürültü azaltımı ve süper çözünürlük performansı çok daha ileri düzeye taşınabilir.



Şekil 13: Grace Entegrasyonu Sonucu Test Karesi

Kaynakça

Cheng, Y. et al., “GRACE: Loss-Resilient Real-Time Video through Neural Codecs,” *arXiv preprint arXiv:2305.12333*, 2024.

Cheng, Y. et al., “DeepRS: Deep-learning-based Network-Adaptive FEC for Real-Time Video Communications,” *IEEE ICASSP*, 2020.

Tung, T. T., & Gündüz, D., “DeepWiVe: Deep-Learning-Aided Wireless Video Transmission,” *IEEE Journal on Selected Areas in Communications*, 2022.

Yan, J. et al., “Salsify: Low-Latency Network Video through Tighter Integration between a Video Codec and a Transport Protocol,” *NSDI*, 2018.

Chen, Y. et al., “NeRV: Neural Representations for Videos,” *Advances in Neural Information Processing Systems*, 2021.

Lu, G. et al., “DVC: An End-to-End Deep Video Compression Framework,” *IEEE CVPR*, 2019.

Wu, C.-Y. et al., “Learning a Deep Video Compression Model with Spatial-Temporal Priors,” *IEEE CVPR*, 2018.

Microsoft Research, “MSR-VTT: A Large-Scale Video Description Dataset,” [Online]. Available: <https://www.microsoft.com/en-us/research/project/msr-vtt>