

**DERİN ÖĞRENME TEKNİKLERİNİ KULLANARAK
RGB-D NESNE TANIMA**

**RGB-D OBJECT RECOGNITION
USING DEEP LEARNING TECHNIQUES**

ALİ ÇAĞLAYAN

DOÇ. DR. AHMET BURAK CAN

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin

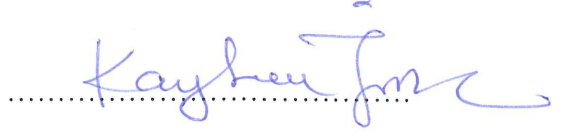
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü

DOKTORA TEZİ olarak hazırlanmıştır.

2018

ALİ ÇAĞLAYAN'ın hazırladığı “Derin Öğrenme Tekniklerini Kullanarak RGB-D Nesne Tanıma” adlı bu çalışma aşağıdaki jüri tarafından BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda DOKTORA TEZİ olarak kabul edilmiştir.

Doç. Dr. Kayhan İMRE
Başkan



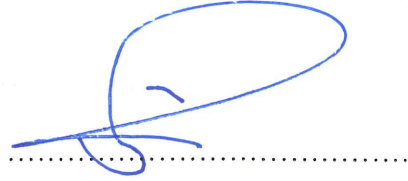
Doç. Dr. Ahmet Burak CAN
Danışman



Doç. Dr. Nazlı İKİZLER CİNBİŞ
Üye



Dr. Öğr. Üyesi Murat ÖZBAYOĞLU
Üye



Dr. Öğr. Üyesi Mustafa SERT
Üye



Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından DOKTORA TEZİ olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin / raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan “ **Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge**” kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. ⁽¹⁾
- o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren Ay ertelenmiştir. ⁽²⁾
- o Tezimle ilgili gizlilik kararı verilmiştir. ⁽³⁾

06/12/2018



Ali ÇAĞLAYAN

“Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge”

- (1) Madde 6. 1. Lisansüstü teze ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir
- (2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. Şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü ve fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.
- (3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.
Madde 7. 2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

* Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.

Aileme...

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

28 / 11 / 2018



Ali ÇAĞLAYAN

ÖZET

DERİN ÖĞRENME TEKNİKLERİNİ KULLANARAK RGB-D NESNE TANIMA

Ali ÇAĞLAYAN

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Doç. Dr. Ahmet Burak CAN

Kasım 2018, 100 sayfa

Nesne tanıma, bilgisayarlı görü alanının temel ve zorlu problemlerinden birisidir. RGB görüntüleri ile beraber zengin geometrik yapıları derinlik verilerini sağlayan Microsoft Kinect gibi algılayıcıların yaygınlaşmalarıyla birlikte, RGB-D verileri, temel bilgisayarlı görü problemlerini çözmek için çok yararlı bir kaynak olarak ortaya çıkmıştır. Özellikle robotik görme alanında bu tür verilerin kullanıldığı nesne tanıma görevi, robotun ortama etkileşiminde ve görsel kavrayışında önemli bir rol oynamaktadır. Öte yandan, derin öğrenme tekniklerinde kaydedilen özellikle son on yıldaki gelişmeler, nesne tanıma performansında büyük bir artış sağlamıştır.

Bu tez kapsamında, derin öğrenme tekniklerini kullanarak RGB-D nesne kategorilerini tanımak için gerçekleştirilen çeşitli çalışmalar sunulmaktadır. Bu çalışmalarda, derin öğrenme tekniklerinden evrimsel sinir ağları (ESA, *convolutional neural networks*) ve özyinelemeli sinir ağları (ÖSA, *recursive neural networks*) kullanılmaktadır. Tezin ilk aşamasında, evrim filtrelerinin gözetimsiz bir şekilde öğrenildiği bir ESA ve bir de ÖSA olmak üzere iki katmanlı, sığ bir mimari kullanılarak RGB-D nesne tanıma için bir analiz çalışması sunulmaktadır. RGB ve derinlik verilerinin farklı karakteristiklerine uygun

olarak, geriyayılım algoritması kullanmaksızın ileri-beslemeli öğrenme gerçekleştiren sığ mimaride, etkin model ayarlamaları ve parametreleri araştırılmaktadır. Tezin sonraki aşamasında, derinlik verilerinde saklı olan zengin geometrik bilgilerden daha iyi faydalanmak için çeşitli hacimsel gösterimler tanımlanarak, bu hacimsel gösterimleri giriş olarak ele alan 3-boyutlu ESA mimarileri ile tanıma gerçekleştirilmektedir. Bu amaçla, derinlik verileri 3B voksel grid temsilleri ile ifade edilmekte ve bu temsillere uygun 3B ESA modelleri deneysel olarak araştırılarak uygun bir model sunulmaktadır. Tezin son kesiminde ise transfer öğrenme ile RGB-D nesne tanıma için yeni bir yaklaşım sunulmaktadır. Buna göre ilk önce bir öneğitilmiş ESA modeli ile RGB ve derinlik verileri için farklı katmanlardan nitelikler çıkartılmaktadır. Daha sonra bu nitelikler daha yüksek düzeyli temsillere eşlenmek üzere, ÖSA modelleri ile dönüştürülmektedir. Son olarak farklı düzeyden çıkartılan temsiller birleştirilerek bir nesne görüntüsünün bütünü ifade eden vektörler elde edilmektedir.

Önerilen çalışmalar, RGB-D nesne tanıma için literatürde sıkça kullanılan veri kümelerinde gerçekleştirilen kapsamlı testler ile analiz edilmektedir. Önerilen yöntemlerde, çalışma amaçlarını doğrulayan ve ilgili çalışmalarla yarışabilir düzeyde, başarılı sonuçlar elde edilmektedir.

Anahtar Kelimeler: RGB-D Nesne Tanıma, Derin Öğrenme, Evrimsel Sinir Ağları, Özyinelemeli Sinir Ağları, Transfer Öğrenme, Derin Nitelikler, Nitelik Öğrenme, Sınıflandırma.

ABSTRACT

RGB-D OBJECT RECOGNITION USING DEEP LEARNING TECHNIQUES

Ali ÇAĞLAYAN

Doctor of Philosophy, Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Ahmet Burak CAN

November 2018, 100 pages

Object recognition is one of the basic and challenging problems of computer vision. With the widespread use of RGB-D sensors such as Microsoft Kinect, which provides rich geometric structured depth data along with RGB images, RGB-D data have emerged as very useful resources for solving fundamental computer vision problems. Particularly in the field of robotic vision, an object recognition task using such data plays an essential role in the interaction of a robot with its surrounding environment and the capability of its visual comprehension. On the other hand, the tremendous progress in deep learning techniques over the last decade, has led to a significant increase in object recognition performance.

In this thesis, several studies on RGB-D object category recognition using deep learning techniques are presented. In these studies, convolutional neural networks (CNN) and recursive neural networks (RNN) are employed. In the first phase of the thesis, an empirical analysis for RGB-D object recognition based on a two-layered shallow architecture with an RNN layer and a CNN layer in which the convolution filters are learned in an unsupervised manner is presented. In accordance with the different

characteristics of RGB and depth data, effective model settings and parameters are investigated in this shallow model that learns deep features in a feed-forward manner without backpropagation algorithm. In the next phase of the thesis, various volumetric representations are defined in order to make better use of the rich geometric information stored in the depth data and recognition is carried out with 3-dimensional CNN architectures that take these volumetric representations as inputs. To this end, depth data are represented by 3D voxel grid representations and a suitable 3D CNN model is presented for these representations by experimentally investigating among many different alternatives. In the last part of the thesis, a new approach based on transfer learning for RGB-D object recognition is presented. To this end, firstly, a pretrained CNN model is used to extract features from different layers for RGB and depth data. Then, these features are transformed with RNN structures to map to higher-level representations. Finally, the representations derived from different levels are fused to produce a final vector expressing the holistic object image.

The proposed works are analyzed with extensive experiments performed on the well-known datasets for RGB-D object recognition. The proposed works produce successful results that confirm the main objectives and the results are highly competitive with the related studies.

Keywords: RGB-D Object Recognition, Deep Learning, Convolutional Neural Networks, Recursive Neural Networks, Transfer Learning, Deep Features, Feature Learning, Classification.

TEŞEKKÜR

Lisans ve lisansüstü eğitim hayatımda çok büyük katkıları olan, tez çalışmam süresince bilgilerinden ve tecrübelerinden faydalanmamı sağlayan, güven ve desteğini hiçbir zaman esirgemeyen, tez danışmanım Sayın Doç. Dr. Ahmet Burak CAN'a,

Tez izleme aşamasında fikir ve önerileriyle tezin gelişmesine katkıda bulunan, tez metnini inceleyerek biçim ve içerik bakımından son halini almasına yardımcı olan, tez izleme komitesi üyesi hocalarım Sayın Doç. Dr. Kayhan İMRE'ye ve Sayın Dr. Öğr. Üyesi Murat ÖZBAYOĞLU'na,

Tez metnini inceleyerek içerik bakımından son halini almasına yardımcı olan, tez savunma sınavım sırasında önerileriyle bana katkıda bulunan Sayın Doç. Dr. Nazlı İKİZLER CİNBİŞ'e ve Sayın Dr. Öğr. Üyesi Mustafa SERT'e,

Tez çalışmalarımın olanaklarından yararlandığım Hacettepe Üniversitesi Bilgisayarlı Görü Laboratuvarı'nın değerli üyelerine,

Tez çalışmalarımın bana her fırsatta yardımcı olan, desteklerini esirgemeyen başta değerli arkadaşlarım Ali Osman SERHATOĞLU, Dr. Ali Seydi KEÇELİ, Dr. Aydın KAYA, Dr. Oğuzhan GÜÇLÜ ve Dr. Tuğba ERDOĞAN olmak üzere tüm çalışma arkadaşlarıma,

Doktora eğitimim süresince fikirlerini ve manevi desteklerini esirgemeyen sevgili arkadaşlarım Burak ÇOPUR, Furkan ÇOPUR, Mahmut KILIÇ ve Murat ÖZKAN'a,

Tüm eğitim hayatım boyunca, bilgilerinden ve tecrübelerinden her fırsatta yararlandığım, güven ve desteğini hiçbir zaman esirgemeyen ve beni akademik çalışmaya teşvik eden, sevgili dayım Mustafa KURTARAN'a,

Benim için hiçbir fedakârlıktan kaçınmayan, her daim yanımda olan, sonsuz sevgi ve desteklerini bana sürekli hissettiren, sevgili annem ve babam başta olmak üzere aileme,

Canı gönülden teşekkür ederim.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	i
ABSTRACT	iii
TEŞEKKÜR	v
İÇİNDEKİLER.....	vi
ŞEKİLLER	ix
ÇİZELGELER.....	xii
TERİMLER VE KISALTMALAR	xiii
1. GİRİŞ.....	1
1.1. Hedef ve Katkılar	4
1.2. Tez Metninin Organizasyonu	5
2. TEORİK BİLGİ.....	8
2.1. Evrişimsel Sinir Ağları (ESA).....	8
2.1.1. Giriş.....	8
2.1.2. ESA Mimarisi	9
2.1.2.1. Evrişim Katmanı.....	9
2.1.2.2. Havuzlama Katmanı	10
2.1.2.3. Tam-bağlantılı Katman.....	10
2.1.3. ESA'nın Eğitilmesi	11
2.1.4. Sonuç.....	13
2.2. Özyinelemeli Sinir Ağları (ÖSA).....	13
2.2.1. Giriş.....	13
2.2.2. ÖSA Mimarisi.....	15
2.2.3. Sonuç.....	17
3. RGB-D NESNE TANIMA İÇİN DERİN NİTELİKLERİN ÖĞRENİLMESİNİN DENEYSEL BİR ANALİZİ.....	18
3.1. Giriş.....	18

3.2. İlgili Çalışmalar	19
3.3. Öğrenme Yöntemi	20
3.3.1. Filtre Öğrenme Modülü	20
3.3.2. ESA Katmanı	21
3.3.2.1. Doğrultucu Birimi	21
3.3.2.2. Havuzlama.....	22
3.3.3. ÖSA Katmanı.....	22
3.3.4. Sınıflandırma.....	23
3.4. Deneysel Değerlendirmeler	24
3.4.1. Washinton RGB-D Nesne Veri Kümesi ve Kullanımı	24
3.4.2. Filtre Öğrenme Yaklaşımlarının Etkileri	26
3.4.3. Doğrultucu Birimlerinin Etkileri.....	28
3.4.4. Havuzlama Yöntemlerinin Etkileri	28
3.4.5. Sınıflandırıcı Karşılaştırmaları.....	29
3.4.6. Tartışmalar	30
3.5. Sonuç	32
4. DERİNLİK VERİLERİNDE 3B ESA KULLANARAK HACİMSEL NESNE TANIMA	
33	
4.1. Giriş	33
4.2. İlgili Çalışmalar	35
4.2.1. El Yapımı Nitelik Tabanlı Yöntemler.....	35
4.2.2. 2.5B ESA Tabanlı Yöntemler	36
4.2.3. 3B ESA Tabanlı Yöntemler	37
4.3. Önerilen Yöntem	37
4.3.1. Hacimsel Temsiller	38
4.3.1.1. İkili Grid (Binary Grid).....	38
4.3.1.2. Yoğunluk Gridi (Intensity Grid).....	39
4.3.2. 3B ESA	40

4.3.3. Çoklu-DönüŖlü YaklaŖım.....	43
4.4. Deneysel Deęerlendirmeler	45
4.4.1. Veri Kümeleri ve Kurulumları.....	45
4.4.1.1. Washington RGB-D Veri Kümesi.....	45
4.4.1.2. 2D3D Nesne Veri Kümesi.....	47
4.4.1.3. Çoklu-DönüŖlü Kurulumu	47
4.4.2. Sonular	48
4.4.2.1. Hacimsel Grid Farklılıkları.....	48
4.4.2.2. Washington RGB-D Nesne Veri Kümesinde KarŖılaŖtırma.....	50
4.4.2.3. 2D3D Nesne Veri Kümesinde KarŖılaŖtırma.....	54
4.4.3. Çoklayan-DönüŖlü YaklaŖımı ve Deneysel Sonular	56
4.4.4. Renk Bilgisinin Kullanıldıęı YaklaŖımlar ve Deneysel Sonular	59
4.5. Sonu	61
5. ÖNEęİTİMLİ BİR ESA MODELİNİ ÇOKLU ÖSA YAPILARI İLE BİRLİKTE KULLANARAK RGB-D NESNE TANIMA	63
5.1. GiriŖ	63
5.2. İlgili alıŖmalar	65
5.3. Önerilen Yöntem	67
5.4. Deneysel Deęerlendirmeler	70
5.4.1. Model Analizi	70
5.4.2. KarŖılaŖtırmalı Sonular.....	75
5.4.3. Hacimsel Tanımda ÖSA Modelinin Kullanımı ve Deneysel Sonular	82
5.5. Sonu	83
6. SONULAR	85
6.1. TartıŖmalar ve Gelecek alıŖmalar.....	88
KAYNAKLAR.....	91
ÖZGEÇMİŖ.....	99

ŞEKİLLER

Sayfa

Şekil 1.1. Kinect kamerasındaki algılayıcı yerleşimleri ve bu algılayıcılardan elde edilen örnek görüntüler.	3
Şekil 2.1. Özyinelemeli sinir ağları, tekrarlayan sinir ağlarının dizilimini ağaç yapısı içerisinde genelleştiren bir yapıya sahiptir. Örnekte görüntünün aşağıdan yukarıya doğru parça-bütün hiyerarşisi görülmektedir. Şekil, [52] çalışmasından uyarlanmıştır.....	14
Şekil 2.2. $K \times 4 \times 4$ ve $K \times 2 \times 2$ boyutlarındaki blokları, giriş çocuk vektörleri olarak ele alan ve her düğümde aynı ağ modelini tekrarlı olarak bir üst vektör olan ana vektörü hesaplamak için kullanan örnek bir özyinelemeli sinir ağı modeli [56].....	16
Şekil 3.1. Öğrenme yönteminin şematik gösterimi.	23
Şekil 3.2. Dört farklı kategori için RGB-D veri kümesindeki nesnelerin ağaç yapısında gösterimi.	25
Şekil 3.3. RGB-D veri kümesindeki farklı nesne örnekleri.....	25
Şekil 3.4. Farklı filtre öğrenme yaklaşımlarının nesne kategorilerini tanımadaki etkisi.....	27
Şekil 3.5. Rastgele giriş görüntülerinden öğrenilen filtreler ile tüm alt kategori nesnelerinden öğrenilen filtrelerin karşılaştırılması.	27
Şekil 3.6. Farklı doğrultucuların etkisi.	28
Şekil 3.7. Havuzlama yöntemlerinin etkisi.....	29
Şekil 3.8. Farklı sınıflandırıcılar arasındaki doğruluk başarısı karşılaştırması.	30
Şekil 3.9. Uygun parametrelerinin kullanıldığı modelin başlangıç modeli ile karşılaştırması.....	31
Şekil 4.1. Elma nesne kategorisine ait örnek hacimsel gösterimler.	40
Şekil 4.2. Önerilen yaklaşımın ağ mimarisi	42
Şekil 4.3. Önerilen çoklu-dönüştürme nesne tanıma yöntemi.....	44
Şekil 4.4. Washington RGB-D Nesne veri kümesinin sınıf-içi çeşitliliğini gösteren alt kategori örneklerinden görüntü örnekleri.....	46
Şekil 4.5. Washington RGB-D Nesne veri kümesinde sınıflararası benzerliği gösteren örnekler.....	46

Şekil 4.6. Hacimsel gridlerin çeşitli parametrelerle Washington RGB-D Nesne veri kümesinde doğrulama kümesi üzerinde etkileri.	49
Şekil 4.7. Hacimsel gridlerin çeşitli parametrelerle Washington RGB-D Nesne veri kümesi test kümesindeki doğruluk performansları.	50
Şekil 4.8. Washington RGB-D Nesne veri kümesindeki her bir kategori için f-skorları. ...	52
Şekil 4.9. Washington RGB-D Nesne veri kümesindeki yanlış sınıflandırılmış kategori örnekleri.	53
Şekil 4.10. Tek-dönüştü yaklaşımın, 2D3D Nesne veri kümesindeki hata matrisi.	55
Şekil 4.11. 2D3D Nesne veri kümesinde, çoklu-dönüştü yaklaşımının hata matrisi.	55
Şekil 4.12. Çoklayan-dönüştü yaklaşımla nesne tanıma.	56
Şekil 4.13. Çoklayan-dönüştü yaklaşımın Washington RGB-D Nesne veri kümesindeki bir bölmede yitim-iterasyon değişimini gösteren eğitim raporu.	57
Şekil 4.14. Çoklayan-dönüştü yaklaşımın Washington RGB-D Nesne veri kümesindeki bir bölmede doğruluk-iterasyon değişimini gösteren eğitim raporu.	58
Şekil 4.15. Renk bilgisinin kullanıldığı yaklaşımların ve tek-dönüştü yaklaşımının Washington RGB-D Nesne veri kümesindeki 2 eğitim/test bölmesindeki ortalama doğruluk başarı oranları.	61
Şekil 5.1. Önerilen yaklaşımın genel görünümü.	67
Şekil 5.2. Özyinelemeli sinir ağlarında kullanılan doğrusal olmayan fonksiyonlar.	71
Şekil 5.3. ÖSA için farklı ezme işlevlerinin sınıflandırma doğruluğu açısından etkileri.	72
Şekil 5.4. ÖSA'nın doğruluk performansı ve nitelik boyutları açısından öneğitimli ESA'dan elde edilen orta katman ham niteliklerindeki etkileri.	73
Şekil 5.5. Önerilen yaklaşımın tekil katmanlar için elde ettiği doğruluk performansı.	74
Şekil 5.6. Önerilen yaklaşımın Washington RGB-D Nesne veri kümesindeki kategoriler için tekil sınıflandırma başarıları.	79
Şekil 5.7. Önerilen yaklaşımın derinlik verilerinde tanıma hata matrisi.	80
Şekil 5.8. Önerilen yaklaşımın RGB verilerinde tanıma hata matrisi.	80
Şekil 5.9. Önerilen yaklaşımın RGB-D verilerinde tanıma hata matrisi.	81

Şekil 5.10. Sıkça karıştırılan nesne kategorilerine ilişkin örnek görüntüler.....	81
Şekil 5.11. Hacimsel verilerde 3B ESA-ÖSA işbirliğinin kullanımı.	82
Şekil 5.12. Hacimsel grid verisi için 3B ESA katmanlarından elde edilen başarı oranlarının, ÖSA uygulanarak elde edilen başarı oranları ile karşılaştırması.....	83
Şekil 6.1. Önerilen yaklaşımların, Washington RGB-D Nesne veri kümesinde derinlik verilerini kullanarak, tekil kategoriler için elde edilen sınıflandırma başarıları.....	90

ÇİZELGELER

Sayfa

Çizelge 3.1. Optimum model parametreleri ile ayarlanmış öğrenme modelinin ilgili çalışmalarla olan karşılaştırması.....	31
Çizelge 4.1. Kullanılan 3B ESA mimarisinin ayrıntıları.....	43
Çizelge 4.2. Washington RGB-D Nesne veri kümesinde derinlik verilerini kullanan ilgili çalışmaların kategori tanıma doğruluğu karşılaştırması.....	51
Çizelge 4.3. Derinlik verilerini kullanan ilgili çalışmaların 2D3D Nesne veri kümesinde kategori tanıma doğruluğu karşılaştırması.....	54
Çizelge 4.4. Önerilen tek-dönüslü, çoklu-dönüslü ve çoklayan-dönüslü yaklaşımlarının Washington RGB-D Nesne ve 2D3D Nesne veri kümelerinde elde ettikleri doğruluk sonuçları.....	58
Çizelge 4.5. 8-bitlik renk kodlanması.....	59
Çizelge 5.1. Orta seviye katmanlarının birleştirilmeleri ile elde edilen farklı kombinasyonlar için RGB ve derinlik verilerinde elde edilen doğruluk performansları	74
Çizelge 5.2. Nihai RGB-D tanıma doğruluk performansı için, RGB ve derinlik verilerinin birlikte değerlendirilmeleri ile elde edilen sonuçlar	75
Çizelge 5.3. Önerilen yaklaşımın, Washington RGB-D Nesne veri kümesindeki ilgili diğer yöntemlerle doğruluk oranı karşılaştırması	77

TERİMLER VE KISALTMALAR

Terimler

3D	Üç boyutlu, 3B
3D CNN	3B ESA
Accuracy	Doğruluk
Activation	Aktivasyon
Activation map	Aktivasyon haritası
Average pooling	Ortalama havuzlama
Background	Arkaplan
Backpropagation	Geriyayılım
Bag-of-words	Sözcük torbası
Batch	Yığın
Bias (n)	Yanlılık
Bias (v)	Sapmak
Binary grid	İkili grid
Compact	Kompakt
Confusion matrix	Hata matrisi
Converge	Yakınsamak
Convolution	Evrişim
Convolutional neural networks	Evrişimsel sinir ağları
Cost function	Maliyet fonksiyonu
Cropping	Kırpma
Dataset	Veri kümesi
Depth map	Derinlik haritası
Descriptor	Tanımlayıcı
Detection	Tespit etme
Domain	Alan
Dropout	Seyreltme
Earlier layers	Daha erken katmanlar
Embedded	Saklı
Feature	Nitelik
Feature detector	Nitelik bulucu
Feature map	Nitelik haritası

Feed-forward	İleri-beslemeli
Fine-tuning	İnce-ayarlama
Forward propagation	İleri yayılım
Fully-connected	Tam-bağlantılı
Graph	Çizge
Grayscale image	Gri tonlamalı görüntü
Ground-truth	Gerçek-referans
Hand-crafted features	El yapımı nitelikler
Hidden markov model	Saklı markov modeli
Hybrid	Hibrit
Hypercube	Hiperküp
Individual	Tekil
Instance	Örneklem
Intensity grid	Yoğunluk gridi
Inter-class similarity	Sınıflar-arası benzerlik
Intra-class variation	Sınıf-içi çeşitlilik
Kernel	Çekirdek
k-means	k-ortalama
Learning rate	Öğrenme oranı
Linear support vector machines	Doğrusal destek vektör makineleri
Loss function	Yitim fonksiyonu
Max pooling	Maksimum havuzlama
Multi-rotational	Çoklu-dönüştürme
Nearest neighbor	En yakın komşu
Non-linearity	Doğrusalsızlık
Normalize	Normalleştirmek
Overfitting	Aşırıuyumlama
Parent	Ata
Parse tree	Ayrıştırma ağacı
Patch	Yama
Point cloud	Nokta bulutu
Pooling	Havuzlama
Postprocessing	Sonışleme
Preprocessing	Önişleme

Pretrained	Öneđitimli
Raw	Ham
Receptive field	Alıcı alan
Rectifier unit	Dođrultucu birim
Recurrent neural networks	Tekrarlayan sinir ađları
Recursive	Özyineleme
Recursive neural networks	Özyinelemeli sinir ađları
Regularization	Düzenleyici
Resize	Yeniden boyutlandırmak
Robust	Gürbüz
Rotation	Dönme
Rotation invariance	Dönme deđişmezliđi
Rotational	Dönüşlü
Segmentation	Bölütleme
Semantic	Semantik
Semi-supervised learning	Yarı-denetimli öğrenme
Single-rotational	Tek-dönüşlü
Sliding window	Kayan pencere
Smoothed curve	Yuvarlatılmış eğri
Spatial pyramid pooling	Uzamsal piramit havuzu
Spatial structure	Uzamsal yapı
Split	Bölme
Squash function	Ezme işlevi
Stochastic gradient descent	Stokastik gradyan düşüş
Stride	(adım) Atlama
Supervised learning	Gözetimli öğrenme
The state-of-the-art	En gelişkin
Transformation	Dönüşüm
Translation	Öteleme
Translational invariant features	Ötelemede deđişmez nitelikler
Turntable	Döner platform
Underfitting	Yetersiz uyumlama
Unsupervised learning	Gözetimsiz öğrenme
Validation	Dođrulama

Validation set accuracy	Doğrulama kümesi doğruluğu
Vanishing gradient problem	Gradyanların kaybolması problemi
Volumetric	Hacimsel
Volumetric representation	Hacimsel temsil
Voxel	Voksel, Hacimsel piksel
Warping	Eğrilme
Whitening	Beyazlatma

Kısaltmalar

3B ESA	3-boyutlu Evrişimsel Sinir Ağları
ESA	Evrişimsel Sinir Ağları
NB	Naïve Bayes
NN	Nearest Neighbor
ÖSA	Özyinelemeli Sinir Ağları
SGD	Stokastik Gradyan Düşüş
SMM	Saklı Markov Modeli
SPM	Spatial Pyramid Pooling
SVM	Support Vector Machine
TSA	Tekrarlayan Sinir Ağları

1. GİRİŞ

Nesne tanıma, modern bilgisayarlı görü alanındaki temel problemlerden birisi olup farklı alanlarda çeşitli uygulamalar ile sıkça kullanılmakta ve sürekli gelişim göstermektedir. Tanıma sistemleri biyomedikal alanında, kromozom tanıma, parmak izi tanıma, iris tanıma, anatomik görüntülerde belli örüntüleri tanıma; gıda alanında tarım ürünlerinde deformasyon ve hastalık tespitinde, tür tanıma uygulamalarında; optik karakter tanıma, el yazısı tanıma, plaka tanıma sistemleri ve trafik gözetleme gibi çeşitli alanlarda geniş bir uygulama alanı bulmaktadır [1]. İnsan beyni, görsel ve derinlik algılama kapasitesi, ses farkındalığı, dokunmaya hassas deri sensörleri ve tat duyuları gibi farklı mekanizmaları işbirliği içerisinde kullanabilen çok karmaşık bir yapıdadır. Öte yandan, bilgisayarlı sistemler hem verileri işleme kapasiteleri hem de bu verilerden anlamlı çıktılar elde etmeleri anlamında çok daha geridedir. David Marr, 1960'lı yıllarda başlayan modern tanıma yönelik çalışmalarda araştırmacıların tanımanın zorluğunu kavrayamadığını ve bunun nedeni olarak da insanın kendisinin tanıma çok iyi olmasını sağlayan mükemmel bir tanıma sistemine sahip olduğunu ileri sürmektedir [2]. Nesne tanıma, günümüzde halen devam etmekte olan temel görsel tanıma zorluklarını içermektedir. Bu zorluklar aşağıdaki gibi özetlenebilir:

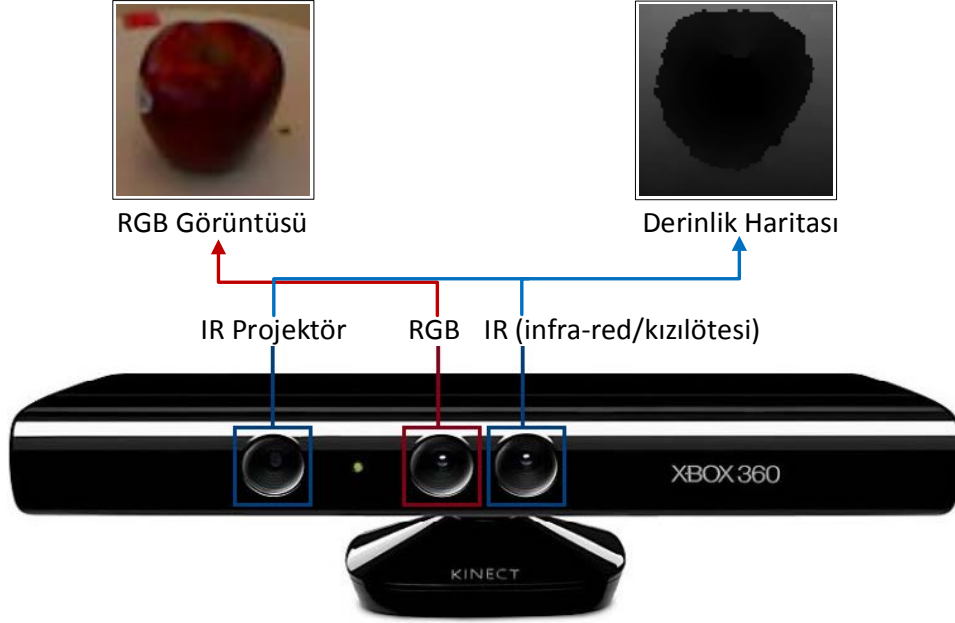
- Tipik bir görüntü pek çok farklı yapılandırmada düzenlenen birçok nesne içerebilir.
- Görüntüler karmaşık ve son derece muğlak olabilirler.
- Nesne kategorilerindeki sınıf-içi çeşitlilik ve sınıflar-arası benzerlik yanlış sınıflandırmaya neden olabilir.
- Görüntü arka planındaki karmaşıklık, nesnelerin ayırt edilmesini zorlaştırabilir.
- Ortamın ışık yapısı, ışığın farklı yüzeylerde yansıtılması ve kırılmasına bağlı doğal zorluklar olabilir.
- Nesnelerdeki bakış açısı ve ölçek değişiklikleri tanımayı zorlaştırabilir.
- Algılayıcı kaynaklı görüntüdeki bozulmalar ve gürültüler tanımayı zorlaştırabilir.

Yukarıda maddelenen tüm bu zorluklarla birlikte nesne tanıma modellerinin tanıma performansını değerlendirme şekli tartışılan bir başka sorundur [3], [4]. Çünkü pratikte değerlendirme için kullanılan veri kümelerindeki görüntülerin çoğunluğuna insan karar vermektedir ve dolayısıyla bu veri kümeleri gerçek dünyadaki çeşitliliği düzgün bir şekilde yansıtmayabilir [3]. İnsanlardaki ve diğer primatlardaki görsel korteks sistemi, neredeyse hiç çaba göstermeden tüm bu zorlukları aşarak nesnelerin tanınmasını inanılmaz bir

doğrulukla gerçek zamanlı olarak gerçekleştirebilmektedir. Biyolojik görsel veri işleme mekanizması, erken kesimleri yönelim ve kenar gibi basit bilgileri algılayan hücreler ile verinin işlendiği yoldan ilerledikçe daha bütüncül nesne kesimlerini algılayan komplike hücreleri içeren, hiyerarşik bir yapıdadır [5], [6]. Biyolojik mekanizmadaki etkinliği ve başarımı elde etmek, modern bilgisayarlı görü alanının nihai hedefi olup, bu kapsamda erken dönemden bu yana analogik çalışmalar yapılmaktadır (örn. Marr Vision [7]). Bu iddialı amaç kapsamında son yıllarda özellikle “derin öğrenme” [8] ile beraber bazı özel görevlerde istenen başarılı sonuçlar elde edilmiştir. Bu tez kapsamında, derin öğrenme yöntemlerini kullanarak RGB-D algılayıcı tabanlı nesne tanıma çalışmaları gerçekleştirilmiştir.

Nesne tanıma problemi, bilgisayarlı görü alanında üzerinde sıkça çalışılan konulardan biri olmuştur. Microsoft Kinect gibi RGB-D algılayıcıların piyasaya sürülmesi ile beraber RGB-D görüntüler, görsel (RGB) görüntülere ek olarak derinlik (D) görüntülerinin tamamlayıcı niteliklerinden ötürü bilgisayarlı görü, bilgisayar grafiği, robotik ve insan-bilgisayar etkileşimi gibi alanlarda yaygın bir şekilde kullanılmaya başlanmıştır. Bu algılayıcıların geometrik veriyi görsel verilerle birlikte eş zamanlı sağlaması, hem akademik alanda hem de endüstride başta oyun sektörü olmak üzere çeşitli alanlarda kullanımlarını yaygınlaştırmıştır. Dolayısıyla bu algılayıcılardan elde edilen görüntülerde her bir piksel, R (*Red-Kırmızı*), G (*Green-Yeşil*) ve B (*Blue-Mavi*) olmak üzere renk değerleri ile D (*Depth-Derinlik*) değerini sağlayan 4-boyutlu [R, G, B, D] vektörü ile temsil edilmektedir. RGB algılayıcı, 640×480 (versiyon 1/ v1) ve 1920×1080 (versiyon 2/ v2) çözünürlüklerinde saniyede 30 çerçeve (30 fps) sağlar. Derinlik algılayıcı ise saniyede 30 çerçeve ile çözünürlükleri 640×480 (v1) ve 512×424 (v2) görüntüler sağlamaktadır. Ancak derinlik algılayıcı teknolojileri v1 ve v2 sürümlerinde farklıdır. Kinect v1, yapılandırılmış ışık (*structured-light*) denilen kızıl ötesi kamera ve kızıl ötesi lazer kaynağı arasında veri üçgenlemesine dayanırken; Kinect v2, *Time-of-Flight* (ToF) denilen bir aydınlatma ünitesinin yaydığı ışığın bir cisme gidip algılayıcıya geri gelmesine kadarki süreyi ölçen teknolojiye dayanmaktadır [9]. Dolayısıyla, Kinect v1’de derinlik algılayıcısının menzili 1.2 ~ 3.5 m iken, Kinect v2’de bu konuda biraz daha iyileştirme yapılarak menzil değerleri 0.5 ~ 4.5 m aralığına çekilmiştir. Microsoft Kinect donanımına ilişkin bir görsel ve RGB ile derinlik algılayıcılarından elde edilen görüntü örnekleri Şekil-1.1’de gösterilmektedir. RGB görüntüleri genellikle nesne görünüşleri ve dokularına ilişkin bilgileri sağlarken, derinlik görüntüleri nesnelerin geometrik yapılarına ilişkin bilgileri

daha iyi temsil eder. Ayrıca, derinlik görüntüleri RGB görüntülerine göre renk, aydınlatma ve bakış açısı değişikliklerine göre daha gürbüz bilgiler sağlar [10]. Bu iki veri türünün bir arada kullanılması, özellikle robotik alanında, robotun ortamın geometrik düzeni ve ortamdaki çevre unsurları hakkında insan algı yeteneklerine benzer mantık yürütmesini sağlayabilir.



Şekil 1.1. Kinect kamerasındaki algılayıcı yerleşimleri ve bu algılayıcılardan elde edilen örnek görüntüler.

Bilgisayarlı görü terminolojisinde “tanıma” (*recognition*) ile birlikte sıkça karşılaşılan ve birbirlerine yakın olan diğer kavramlar, “tespit etme” (*detection*), “yerini belirleme” (*localization*) ve “kavrama” (*understanding*) terimleridir. Tespit etme probleminde, belirli bir görüntüdeki nesnenin varlığı ile ilgilenilir. Yerini belirleme probleminde, tespit etme problemi genişletilerek ilgili nesnenin tam olarak nerede olduğu bilgisi de sorgulanmaktadır. Bu iki kavram genellikle beraber ya da aynı amaçla değiştirilebilir bir biçimde kullanılır. Genellikle nesne tarafından kapsanan alan etrafındaki bir sınırlayıcı kutu (*bounding box*) belirlenerek, bu nesnenin görüntüdeki yeri ve büyüklüğünün kaba bir tahmini verilmektedir. Tanıma bir görüntüde algılanan nesnelerin tanımlanması ve doğrulanması işlemi olarak kabul edilmektedir. Bir diğer ifadeyle görüntüdeki nesnelere belirli sınıflara ayırmak olan kategorizasyon problemi olarak da düşünülebilir. Nesne tanıma probleminde, görüntüdeki nesnenin ne olduğu sorusuna cevap aranır. Son olarak kavrama probleminde, görüntüdeki nesnelerin bağlam içerisindeki rolleri de sorgulanmak suretiyle tanıma probleminin kapsamı genişletilmektedir [1].

Nesne tanımı ve ne olduğu bilgisi, biraz belirsiz ve yapılan işe ve hedeflenen sonuca göre değişebilmektedir. Bir nesnenin; (i) uzayda iyi tanımlanmış bir kapalı sınırı (ii) çevrelerinden farklı bir görünümü (iii) görüntü içerisinde benzersiz ve dikkat çekici yapısı olmak üzere en az üç ayırt edici özellikten birine sahip olması gerekir [11]. Birçok nesne aynı anda bu özelliklerin birkaçını içerebilir. Tipik bir nesne tanıma sisteminde, nesneye ait nitelikler çıkartılır, çıkartılan nitelikler gruplandırılır, gruplandırılan niteliklerin ne olduğu hakkında hipotez kuran bir makine öğrenme algoritması kullanılarak hipotezin doğru olup olmadığı test edilir [1]. Yakın zamana kadar sıklıkla kullanılan bu geleneksel yöntemde, anılan bileşenlere ek olarak ön işleme (*preprocessing*) ve son işleme (*postprocessing*) gibi başka adımların eklenmesi de mümkündür. Bir nitelik (*feature*) görüntüde belli bir noktadan elde edilen bilgilendirici bir ipucudur. Görüntünün genel nitelik kümesi ise genellikle görüntü karakteristiğini temsil eden bir vektörle ifade edilir. Nesne tanımda başarı, büyük ölçüde, birbirleri ile uyumlu etkin nitelikleri çıkartmaya bağlıdır. Ancak etkin nitelik çıkartma, iyi bir alan bilgisi gerektiren zorlu bir iştir. Son yıllarda geliştirilen derin öğrenme teknikleri ile alan uzmanları tarafından tasarlanabilecek karmaşıklık düzeyini ve çeşitliliğini aşan, girdi verilerini daha iyi temsil edebilen nitelikler, otomatik olarak öğrenilebilmektedir. Derin öğrenme yöntemlerinin klasik yöntemlerden farklı olarak verinin temel piksel düzeyinden, hiyerarşik, katmanlı bir yapıda bilgi çıkartması, başarımın temel etmenlerinden birisidir.

1.1. Hedef ve Katkılar

Tez çalışmaları kapsamında, Microsoft Kinect gibi bir RGB-D algılayıcıdan elde edilen görüntüler kullanılarak, görsel ve şekilsel ipuçlarını bir arada kullanabilen, nesne kategorilerinin otomatik olarak algılanıp tanındığı yöntemlerin geliştirilmesi hedeflenmiştir. Bu amaçla RGB-D algılayıcıların sağladığı renk (RGB) ve derinlik (D) görüntülerini girdi olarak ele alan, temelinde insan görme sistemine benzer yaklaşımla nesne kategorilerini hiyerarşik bir şekilde öğrenen derin öğrenme teknikleri ile nesnelere tanınmıştır. Yapılan çalışmalar, RGB-D nesne tanıma problemi için yaygın kullanılan veri kümeleri üzerinde literatürdeki ilgili çalışmalarla kıyaslanarak değerlendirilmiştir. Tez kapsamında, derin öğrenme tekniklerinden evrimsel sinir ağları (ESA, *convolutional neural networks*) ve özyinelemeli sinir ağları (ÖSA, *recursive neural networks*) kullanılmıştır. İlk aşamada, geriyayılım (*backpropagation*) algoritması olmaksızın, gözetimsiz filtre öğrenmeye dayalı, bir ESA katmanı bir de ÖSA katmanı olmak üzere iki katmanlı ileri-beslemeli (feed-forward) sığ bir mimari üzerinde RGB-D nesne tanıma

problemi için derin nitelikleri öğrenen, deneysel bir analiz çalışması yapılmıştır. Bu analiz çalışmasına göre RGB ve derinlik görüntüleri için farklı filtre öğrenme, ESA yapısındaki doğrultucu birim (*rectifier unit*) ve havuzlama (*pooling*) yaklaşımları ile çeşitli sınıflandırıcılar denenmiştir. RGB ve derinlik verilerine uygun bulunan model parametreleri ile sistem başarısı artırılmıştır. Bu çalışmadaki yaklaşıma göre, derinlik verileri RGB verilerine ek bir kanal olarak ele alınmaktadır. Dolayısıyla derinlik verilerindeki, saklı zengin geometrik bilgilerini daha iyi ifade etmek için, sonraki bölümde, iki tür hacimsel grid temsili tanımlanmıştır. Bu hacimsel temsilleri girdi olarak kullanan hem tek bir gösterim hem de birden çok gösterimleri beraber kullanabilen bir 3-boyutlu ESA (*3D CNN*) mimarisi ile tanıma gerçekleştirilmiştir. Son olarak, daha büyük veri kümelerinde eğitilmiş modelleri kullanarak başarıyı daha da artırmak için, transfer öğrenme ile RGB-D nesne tanıma gerçekleştirilmiştir. Bu kapsamda, öneğitimli bir ESA (*pretrained CNN*) modeli ile farklı katmanlardan nitelikler çıkartılmış, bu nitelikler girdi olarak tek seviyeli bir ÖSA katmanına verilmiştir. Buradan elde edilen nihai çıktılar çeşitli seviyelerde birleştirilerek sistem başarımı önemli derecede artırılmıştır. Derinlik verilerini öneğitimli ESA modelinde kullanabilmek için, derinlik verilerinden yüzey normalleri hesaplanarak RGB verilerine benzer bir yapıda renklendirme işlemi gerçekleştirilmiştir.

1.2. Tez Metninin Organizasyonu

Tezin bu ilk giriş bölümünde, bu tezde sunulan çalışmaların motivasyon kaynakları, temel hedefleri ve katkıları anlatılmaktadır. Geri kalan bölümler şu şekilde tasarlanmıştır:

- Bölüm 2, tez çalışmalarında konu edilen derin öğrenme tekniklerinden evrimsel ve özyinelemeli sinir ağlarının özet bir teorik altyapı bilgisini kapsayacaktır. Derin öğrenmenin, son yıllarda başta bilgisayarlı görü alanında olmak üzere, makine öğrenmesinin konu edindiği tüm alanlarda gösterdiği üstün performansından ötürü, literatürde ilgili çok çeşitli çalışmalar mevcuttur. Bu bölümde, tez kapsamında yapılan çalışmaların okunurluğunu kolaylaştırmak için temel kavramlar ana hatları ile konu edilecektir.
- Bölüm 3'te, RGB-D nesne tanıma için derin nitelikleri öğrenen analiz çalışması ele alınmıştır. Bu çalışmada kullanılan çatı mimari literatürde yaygın bir şekilde kullanılan bir ESA bir de ÖSA katmanından oluşan sığ bir modeldir. Bu çalışma ile literatürdeki çalışmalardan farklı olarak RGB ve derinlik verileri için uygun model parametreleri araştırılarak, verinin karakteristiğine uygun parametrelerle başarı derecesinin artırılabilirdiği gösterilmiştir. RGB ve derinlik verileri için farklı

havuzlama ve dođrultucu fonksiyonlarının daha iyi sonu verdiđi grlmştr. te yandan gzetimsiz filtre đrenilirken, rastgele noktalardan yama (*patch*) ıkartmanın, yaygın kullanılan nitelik bulucu (*feature detector*) noktaları etrafından ıkartılan yamalardan, pek farkı olmadan aynı derecede bařarı sađladıđı gzlenmiřtir. Bu alıřmanın temel etkisi, gzetimsiz filtre đrenmeye dayalı bu tr sıđ modelleri RGB-D verilerinde kullanırken modelde yapılacak radikal deđiřimler yerine veri trlerine uygun dikkatli parametrelerin seimi ile bařarının nemli oranda artırılabilirdiđidir. Bu blmdeki alıřmalar, [12] bildirisinde sunulmuřtur.

- Evriřimsel sinir ađlarındaki geliřmeler, nitelik tasarımı dikkate almadan niteliklerin otomatik olarak đrenilmesine ve nesne tanımda daha iyi sonuların retilmesine olanak sađlamıřtır. Derinlik ve RGB verilerini kullanan birok ESA alıřmaları nerilmelerine rađmen, derinlik verilerinde gizlenmiř hacimsel bilgiler tam olarak kullanılmamaktadır. Bu alıřmaların ođunda derinlik verisi RGB'ye ek bir kanalmıř gibi ele alınmaktadır. Blm 3'te konu edinen alıřma da bu tarz bir yaklařım ile derinlik verilerini RGB renk kanallarına ek bir kanalmıř gibi ele almaktadır. Oysa bu iki veri trnn karakteristiđi farklıdır. Derinlik verilerini ekstra bir kanalmıř gibi kullanmak yerine nesnelerin geometrik yapısını ortaya ıkartacak gsterimlerle ele almak daha iyi olabilir. Dolayısıyla, bu yaklařımı iyileřtirmek iin derinlik verilerindeki saklı olan geometrik bilgileri daha iyi ifade edebilen hacimsel yaklařımlı nesne tanıma alıřmaları, Blm 4'te sunulmaktadır. Bu blmde anlatılan alıřmalarda, bu amala, derinlik grntlerinde saklı zengin 3B yapısal bilgileri ortaya ıkartmak iin iki tr hacimsel gsterim nerilmektedir. Bu hacimsel gsterimleri giriř olarak alan, tek bir hacimsel gsterimden ve oklu hacimsel gsterimlerden nesnelere tanıyan 3B ESA modellenli farklı yaklařımlar nerilmektedir. Tek bir hacimsel gsterimden nesne tanıma gerekleřtiren yaklařım, yaygın kullanılan iki veri kmesinde ilgili diđer alıřmalarla rekabeti sonular retmektedir. Bununla birlikte, nesnelerin oklu dnřleri bir araya getirildiđinde tanıma dođruluđu daha da artmaktadır. oklu-dnřl 3B ESA yaklařımı, dnme deđiřmezliđini sađlamak iin birden fazla hacimsel gsterimden gelen bilgileri birleřtirerek, tek-dnřl yaklařıma gre bařarı derecesini nemli derecede geliřtirmektedir. Ayrıca, nesnelerin farklı aılardan ekilmiř grntlerini bir araya getiren oklu-dnřl yaklařımına benzer řekilde, giriř olarak alınan hacimsel temsilleri eřitli aılardan dndrmek suretiyle ođaltan oklayan-dnřl yaklařım nerilmektedir. Deneysel sonular, nesnelerin oklu grnmlerinin

kullanılmasının, 3B ESA tabanlı nesne tanıma için oldukça bilgilendirici olabileceğini göstermektedir. Bunların yanı sıra, hacimsel gösterimler içerisinde renk bilgisini kodlayan yaklaşımlar da önerilerek deneysel sonuçlar verilmektedir. Renk bilgisinin çeşitli şekillerde kullanıldığı ESA modellerinde yapılan deneysel analizler sunulmaktadır. Önerilen yaklaşımlar, yaygın kullanılan iki veri kümesindeki ilk hacimsel nesne tanıma çalışmaları olup, literatürdeki diğer ilgili çalışmalarla rekabetçi sonuçlar vermektedir. Bu bölümde anlatılan çalışmalar [13] ve [14] yayınlarında sunulmuştur. Bu tez metninde bu çalışmaların kapsamı genişletilmektedir.

- Bölüm 4'te anlatılan çalışmalarda, 3B ESA modelinde eğitim sıfırdan gerçekleştirilmektedir. Öte yandan, transfer öğrenme ile çok büyük ölçekli veri kümelerinde günlerce eğitilen hazır modelleri kullanarak başarıyı daha da artırmak mümkündür. Makine öğrenme teknikleri, genellikle eğitim ve test verilerinin aynı alanda (*domain*) olduğu veri uzayında öğrenme gerçekleştirilmektedir. Ancak eğitim verilerini hem toplamak hem de modelleri bu veriler üzerinde eğitmek pahalı ve yorucudur. Transfer öğrenme, geleneksel tekniklerden farklı olarak başka alanda ve/veya veri kümelerinde öğrenen modellerin mevcut bir probleme hızlı ve kolay bir şekilde uygulanmasına ya da adapte edilmesine olanak sağlar. Özellikle geniş ölçekli RGB veri kümeleri üzerinde eğitilmiş modellerin, nispeten daha küçük veri kümelerinin mevcut olduğu RGB-D (RGB ve derinlik) alanına uygulanması çok önemli olmaktadır. Bölüm 5, büyük ölçekli veri kümesinde eğitilmiş bir modelden öğrenilmiş nitelikleri, transfer öğrenme yaklaşımına göre ele alan bir yaklaşım sunmaktadır. Bu çalışmada, ilk önce öneğitilmiş bir ESA modeli kullanımı ile alt düzey nitelikler çıkartılarak, bunlara, çoklu ÖSA yapıları uygulanmak suretiyle daha yüksek düzeyli nitelik temsilleri elde edilmektedir. RGB veri kümesi üzerinde öneğitilmiş modeli, derinlik alanında kullanmak için, derinlik verileri RGB'ye benzer bir şekilde kodlanmaktadır. Tek bir seviyeden elde edilen nitelikler yerine farklı seviyelerden nitelikler çıkartılmakta ve çeşitli analizlerle en iyi sonucu veren katman nitelikleri birleştirilerek tanıma işlemi gerçekleştirilmektedir. Bu bölümdeki sonuçlar [15] bildirisinde sunulmuştur. Bu tez metni kapsamında, bu çalışma genişletilmektedir.
- Bölüm 6'da bu tez kapsamında yapılan çalışmalar özetlenip elde edilen sonuçlar kısaca değerlendirilmektedir. Konu edilen çalışmalarla ilgili tartışmalara yer verilip olası gelecek çalışmalar irdelenmektedir.

2. TEORİK BİLGİ

Derin öğrenme, makine öğrenmesinin bir alt kümesi olarak ortaya çıkan, son yıllarda görüntü ve video tabanlı öğrenme, ses tanıma, tavsiye sistemleri ve doğal dil işleme gibi çeşitli alan çalışmalarının büyük bir kısmını domine eden, muhtelif hiyerarşik mimarilerle verilerden yüksek düzeyli soyutlamaları öğrenen bir alandır [16]. Derin öğrenme tekniklerindeki gelişmelerle beraber, el yapımı (*hand-crafted*) nitelik tasarlamaya daha az ihtiyaç duyulmuştur. Çünkü derin öğrenme yöntemleri, veriyi temel düzeyden üst düzey gösterimlere kadar seviye seviye işleterek, hiyerarşik bir şekilde öğrenmeyi gerçekleştirir. Son yıllarda özellikle, birçok derin öğrenme tekniği geliştirilmiştir. Bu bölüm, tez çalışmaları kapsamında ele alınan tekniklerden evrişimsel sinir ağlarına (ESA) ve özyinelemeli sinir ağlarına (ÖSA) esas olan temel kavramlara, yakın bir bakış içermektedir.

2.1. Evrişimsel Sinir Ağları (ESA)

2.1.1. Giriş

Evrişimsel sinir ağları, temelinde matematiksel bir işlem olan evrişim işlemini barındıran ve biyolojik sinir ağlarından esinlenerek yaratılmış olan popüler bir derin öğrenme tekniğidir. 1960'lı yıllarda Hubel ve Wiesel isimli araştırmacılar, yaptıkları çalışmalarla [5], [17] inceledikleri hayvanların görsel korteksindeki hücrelerin, alıcı alan (*receptive field*) olarak adlandırdıkları yapılarındaki ışığı saptamaktan sorumlu olduğunu buldular. Görsel korteksteki karmaşık hücre düzeninin, görsel alanın örtüşen ve küçük alt bölgelerindeki ışıktan sorumlu olduğunu tespit ettiler. Daha sonra bu çalışmalardan esinlenerek, ESA'nın atası sayılan *Neocognitron* [18] diye adlandırılan bir mimarı önerilmiştir. *Neocognitron*'a benzer ancak geriyayılım (*backpropagation*) algoritmasını [19] kullanan ve ABD posta kodlarını tanıyan sistem olan [20] çalışması ile bu sistemin sonraki yıllarda geliştirilen *LeNet-5* [21] olarak adlandırılan versiyonu, ESA kullanan temel çalışmalardır. Yapay sinir ağları çalışmalarında uzun bir duraklama döneminden sonra 2012 yılında, AlexNet [22] olarak bilinen çalışma görsel tanıma alanında yeni bir çığır açtı. Bu başarıda üç temel unsurun rolü önemlidir. Birincisi, bilgisayarların işlem gücü, bilhassa grafik işlemcilerin bu tür çalışmalarda kullanılmasıdır. İkincisi, ImageNet [23] gibi büyük ölçekli veri kümelerinden öğrenmenin gerçekleştirilmesidir. Son olarak önemli gelişmelerden biri de, bu tür mimarilerde etkili bir şekilde öğrenmeyi mümkün kılacak daha iyi bir iklendirme ya da daha iyi bir aktivasyon fonksiyonunun kullanımı gibi mimariye ilişkin bir takım model parametrelerindeki gelişmelerdir. Aslında senaryonun

buna benzer olacağına dair isabetli öngörüler daha önceden yapılmıştı. Örneğin 1992’de yapılmış [24] çalışmasında, yeterli örnek/veri ve yeterli bilgi işlem gücü ile bu tür mimarilerin performanslarının eldeki bir görev için mümkün olan en iyi yaklaşıma sahip olacakları ve önişleme ya da özel temsillere ihtiyaç duyulmaksızın ham veriden öğrenmenin mümkün olacağı belirtilmektedir. Ancak tüm bu şartların ve gelişmelerin ortaya çıkması, 20 yıl sonra mümkün olmuştur.

2.1.2. ESA Mimarisi

Genel olarak ESA, evrişim katmanı (*convolution layer*), havuzlama katmanı (*pooling layer*) ve tam-bağlantılı katman (*fully-connected layer*) olmak üzere 3 temel bileşenden oluşmaktadır.

2.1.2.1. Evrişim Katmanı

Giriş verilerinin uzamsal yapılarını dikkate alan evrişim katmanı, ESA’ların temel bileşenidir. Matematiksel olarak bir evrişim katmanı aşağıdaki gibi ifade edilebilir:

$$h_{l+1} = f(W_l * h_l + b_l) \quad (2.1)$$

Buradaki h_{l+1} , $(l + 1)$. saklı katmandaki çıktı sonucunu, W_l , h_l ve b_l sırasıyla bir önceki katmandaki çekirdek (*kernel*) ağırlıklarını, nitelik haritasını (*feature map*) ve yanlılık (*bias*) değerlerini ifade eder. f ise bir önceki katmandan elde edilen sonuca elementel doğrusalsızlık (*non-linearity*) işlemini uygulayan bir aktivasyon fonksiyonunu temsil eder. Aktivasyon fonksiyonlarının rolü, tüm yapay sinir ağlarında olduğu gibi ESA’da da önemlidir. Çünkü aktivasyon fonksiyonu kullanılmayınca, ağ basit bir şekilde veriyi doğrusal olarak modelleyen bir yapıda olur. Bu nedenle, büyük ve karmaşık veri kümelerinden, temsil kabiliyeti yüksek komplike gösterimler öğrenebilmek ve katman çıktılarını daha güçlü hale getirmek için aktivasyon fonksiyonları kullanılmaktadır. Evrişim katmanlarının kullanım amaçları, erken katmanlarda kenar, köşe ve uç noktaları yakalayıp, bu ipuçlarını seviyeli ve hiyerarşik bir şekilde birleştirip giriş verisini temsil edebilen üst seviyeli gösterimleri öğrenmektir. Bu amaçla, evrişim katmanları; yerel etkileşimler (*local interactions*), paylaşılan ağırlıklar (*shared weights*) ve eşdeğişkenli temsiller (*equivariant representations* veya *equivariance to translations*) olmak üzere üç temel kavramdan yararlanırlar [25]. ESA’da nöronlar, giriş verisinin sadece belirli bir yerel alanına bağlı olmaktadır. Bu yerel alanın uzamsal boyutları, nöronların alıcı alanları (*receptive fields*) ya da filtre/ağırlık boyutları olarak bilinen bir hiper-parametredir. Nöronların alıcı alan kavramı, hem güçlü bir biyolojik ilhama dayalıdır hem de görüntü gibi yüksek boyutlu giriş verilerini ele alırken, modelin bellek gereksinimini azaltıp

istatistiksel verimliliğini artırmaktadır. Giriş verisine nöronların alıcı alanları ve belirli bir adım sayısı (*stride size*) kadar kayırarak filtreler evriştirilmektedir. Bir katmandaki her bir nöron için aynı filtre tüm giriş hacmi (*input volume*) üzerinde gezdirilerek uygulanır. Buna, ağırlıkların paylaşılması kavramı denilmektedir. Böylece örneğin yuvarlak bir kenar bilgisini algılayan bir filtreye, giriş hacminin farklı bölgelerinde verilen tepki ele alınabilmektedir. Zaman serileri verileri işlenirken, bu, evrişimin, giriş verisinde farklı nitelikler görüldüğünde bunları gösteren bir zaman çizelgesi üreteceği anlamına gelir [25]. Paylaşılan ağırlıklar, parametre sayısını düşürerek hesaplama karmaşıklığını azaltırken, ayrıca öteleme değişmezliğine (*translation invariance*) uyum sağlanmaktadır. Evrişim, görüntü ölçeği (*scale*) ve dönme (*rotation*) gibi diğer dönüşümlerdeki (*transformations*) değişikliklere doğal olarak eşdeğer değildir. Bu tür dönüşümleri ele almak için başka mekanizmalar gereklidir [25]. ESA’larda birçok filtre kullanılmaktadır ve her birinin farklı bir niteliği algılaması sezgisel olarak beklenmektedir. Bu filtrelerin bütününe ayrıca filtre bankası denilmektedir.

2.1.2.2. Havuzlama Katmanı

Havuzlama katmanı, girdi olarak ele aldıkları nitelik haritalarından çıkartılan bilgilerin özetini sunarak, hem verinin uzamsal boyutunu ve parametre sayısını, hem de modelin aşırıuyumlamasını (*overfitting*) azaltır. Bunların yanı sıra ayrıca küçük yerel ötelemelere karşı değişmezlik sağlar. Havuzlama işlemi, belli bir çerçeve boyutu boyunca ve belli bir adım sayısı kayırarak giriş hacmine uygulanır. Yaygın olarak kullanılan havuzlama yöntemleri maksimum ve ortalama havuzlama yöntemleridir. Öte yandan, havuzlama işlemi bazı bilgilerin göz ardı edilmesine neden olur. Bundan ötürü yetersiz uyumlamaya (*underfitting*) neden olabilir. Bir görev eğer belirli uzamsal bilgilerin korunmasına dayanıyorsa, tüm nitelik için havuzlama kullanmak modelin eğitim hatasını artırabilir [25]. Örneğin [26] çalışmasında, şekillerin yeniden yapılandırılmaları (*shape reconstruction*) görevinde, havuzlama işlemi belirsizliklere yol açabileceğinden dolayı tercih edilmemektedir.

2.1.2.3. Tam-bağlantılı Katman

Tam-bağlantılı katmanlardaki nöronlar, geleneksel yapay sinir ağlarında olduğu gibi önceki katmandaki tüm nöronlara bağlı olurlar. ESA’da bu katmanların işlevi genellikle üst düzey akıl yürütmektir ve bu katmanlar, ağların son katmanlarını oluşturmaktadırlar. Ancak bu katmanların pratik, işlenebilir olması için boyutlarının makul düzeyde olması gerekir. Son olarak bu katmanların kullanılmadığı, bunlar yerine 1×1 evrişimsel

katmanların tercih edildiği çalışmalar da mevcuttur [27]. Bu tür mimariler, en ve boyu kapsayan uzamsal boyutta bir değişiklik yapmaz iken, giriş olarak alınan hacmin (*input volume*) derinliğini filtre sayısına bağlı olarak ayarlama esnekliğini sağlamaktadırlar.

Literatürde birçok ESA mimari türleri mevcuttur. Ancak temelde, yukarıda özetle bahsedilen katmanları içermektedirler.

2.1.3. ESA'nın Eğitilmesi

ESA'nın eğitilmesi için, sınıflandırılacak verilerin gerçek etiketleri ile ağın tahmin ettiği etiketler arasındaki sapmanın cezalandırılması gerekir. Bu amaçla bir tür yitim fonksiyonu (*loss function*) kullanılmaktadır. Literatürde yaygın kullanılan Softmax yitim, normalleştirilmiş üstel bir fonksiyon olup, k boyutlu bir sınıflandırma görevinde, sınıflar üzerinde $[0 - 1]$ aralığında olasılıksal bir dağılım üretmektedir. Softmax fonksiyonu aşağıdaki gibidir:

$$P(Y_i) = e^{Y_i} / \sum_{j=0}^k e^{Y_j} \quad (2.2)$$

Buradaki Y_i , ESA'da genellikle tam-bağlantılı katmanı takip eden aktivasyon çıktısı olup, $Y_i = w_j^T x + b$ şeklinde ifade edilebilir. Softmax sonucu, belirli bir örneğin belirli bir sınıfa ait olma olasılığını, $[0 - 1]$ aralığındaki bir güven değeri ile vermektedir.

ESA'da giriş olarak etiketli belirli bir veri, çıktı olarak da yitim fonksiyonunun yardımı ile ağın bu etiketli veri için tahmin ettiği sınıf etiketleri bulunur. Bu amaçla, ağ katmanları arasındaki ağırlıkların, başka bir ifadeyle filtrelerin en uygun olanlarını öğrenmek gerekir. Buradan hareketle ESA'nın eğitilmesi, yitim işlevini ya da fonksiyonunu en aza indiren ağ katmanları arasındaki ağırlık kümelerini bulma optimizasyon süreci olarak ifade edilebilir. Bunun için, yitim fonksiyonu başka bir ifadeyle maliyet fonksiyonunu (*cost function*) en aza indiren geri yayılım (*backpropagation*) algoritması kullanılır. Bu amaçla, optimizasyon işlemi olarak yaygın bir şekilde Stokastik Gradyan Düşüş (*SGD, Stochastic Gradient Descent*) algoritması, yığın (*batch*) modda kullanılmaktadır. Yani, eğitim veri kümesini tümünden ele almak yerine kullanılan bilgisayar kaynaklarının el verdiği ölçüde, veri kümesi yığınlara bölünerek ele alınmaktadır. Yığın boyutları 2'nin katları olarak eğitim süreci başlamadan parametre olarak ayarlanır.

Ağı eğitmek için, biri ileri aşama (*forward stage*) diğeri ise geri aşama (*backward stage*) olmak üzere iki aşama vardır. İleri aşamada, giriş görüntüsü mevcut ağırlık değerleri ile

temsil edilir. Daha sonra ağın giriş örnekleri için ürettiği sonucun, olması gereken sonuca olan uzaklığını ölçen maliyet hesabı (*cost function*) yapılır. Geri aşamada ise maliyet hesabına bağlı olarak ağı çıktısı değerinin doğru yönde ayarlanmasını ifade eden her bir parametrenin gradyan hesabı yapılmakta ve tüm parametreler hesaplanan bu gradyan değerlerine göre güncellenmektedir. Bu işlemler, tek bir iterasyonda yapılan adımlardır. Öğrenme, maliyet değeri, makul ve istenen düzeye gelinceye kadar birden çok iterasyonla bu işlemlerin tekrarlanmasıyla tamamlanmaktadır.

SGD, standart gradyan düşüş (*vanilla* ya da *batch GD* olarak da bilinir) algoritmasından farklı olarak, bir iterasyonu tamamlamak için verilerinin tümünü ele almak yerine, verinin küçük bir yığını üzerinde gradyan hesaplaması yapılır. Dolayısıyla algoritmanın yakınsaması (*converge*), hesap maliyeti çok daha az olan fakat daha fazla adımda gerçekleşir. Sonuç olarak daha etkili bir şekilde sonuca ulaşılır. SGD algoritması uygulanmadan önce, veri kümesi rastgele karıştırılır. Böylece alınan alt kümeler veri kümesinin genel dağılımını temsil eder ve karıştırılan veriler, varyansı ve aşırıuyumlamayı azaltıp modellerin genel kalmasına hizmet eder. Ağırlık güncellenmesi ise aşağıdaki gibi yapılmaktadır:

$$W_{t+1} = W_t - \alpha \nabla f_t(W) \quad (2.3)$$

Bu formüle göre $(t + 1)$ iterasyonundaki W ağırlık değerleri t iterasyonu ileri aşamasında hesaplanan gradyan değerlerine göre güncellenir. Buradaki α “öğrenme oranı” ya da “adım sayısı” olarak ifade edilen bir hiper-parametredir. Bu değer uygun bir şekilde seçilmesi etkili bir öğrenme için kritiktir. Çok küçük seçilirse algoritmanın yakınsaması zaman alırken, büyük seçilmesi ise yakınsama noktasının es geçilip öğrenmenin başarısız olmasına neden olabilir. α değerinin seçilmesi yanı sıra W başlangıç ağırlık değerlerinin öğrenmeden önceden ilklendirilmesi gerekir. ESA’lar çok büyük sayıda parametre içermektedirler ve eğitilmeleri oldukça zordur. Bu yüzden eğitim aşamasında, modelin hızlı bir şekilde yakınsaması ve gradyanların kaybolması probleminden (*vanishing gradient problem*) kaçınmak için ağırlık değerlerinin uygun bir şekilde ilklendirilmesi gerekir [28]. Standart sapması 0.01 olan sıfır-ortalama Gauss dağılımı (zero-mean Gaussian distribution) ile ilklendirme [22], giriş/çıkış nöron sayısını dikkate alarak ilklendirme yapan “Xavier” yöntemi [29], ReLU doğrusalsızlığını dikkate alarak Xavier yöntemini oldukça derin modellerde yakınsatabilen He ve diğerleri tarafından önerilen yöntem [30] ve Mishkin ile Matas’ın önerdikleri yöntem [31] literatürde yaygın kullanılan ilklendirme yaklaşımlarıdır.

ESA'ların performansı, birçok derin öğrenme tekniklerinde olduğu gibi geniş ölçekli veri kümeleri üzerinde eğitilmelerine bağlıdır. Bu yüzden eğitim esnasında yaygın kullanılan bir yaklaşım da verinin çoğaltılmasıdır. Bu amaçla, eldeki verilere eklenen rastgele pertürbasyonlarla çoğaltma, döndürme, kaydırma ve aynalama gibi geometrik dönüştürme yöntemleri yaygın bir şekilde kullanılmaktadır. Böylece modelin ağda kodlanması kolay olmayan dönüşümlere karşı değişmez olması sağlanırken, eldeki örnekleri ezberleme eğilimini de ortadan kaldırır. Verinin çoğaltılması teknikleri yanı sıra, Dropout [32], DropConnect [33], $l1$ ile $l2$ ağırlık düzenleyicileri ile elastic net [34] düzenleyicileri gibi yöntemler, modellerin genelleme yeteneklerini oldukça geliştirir.

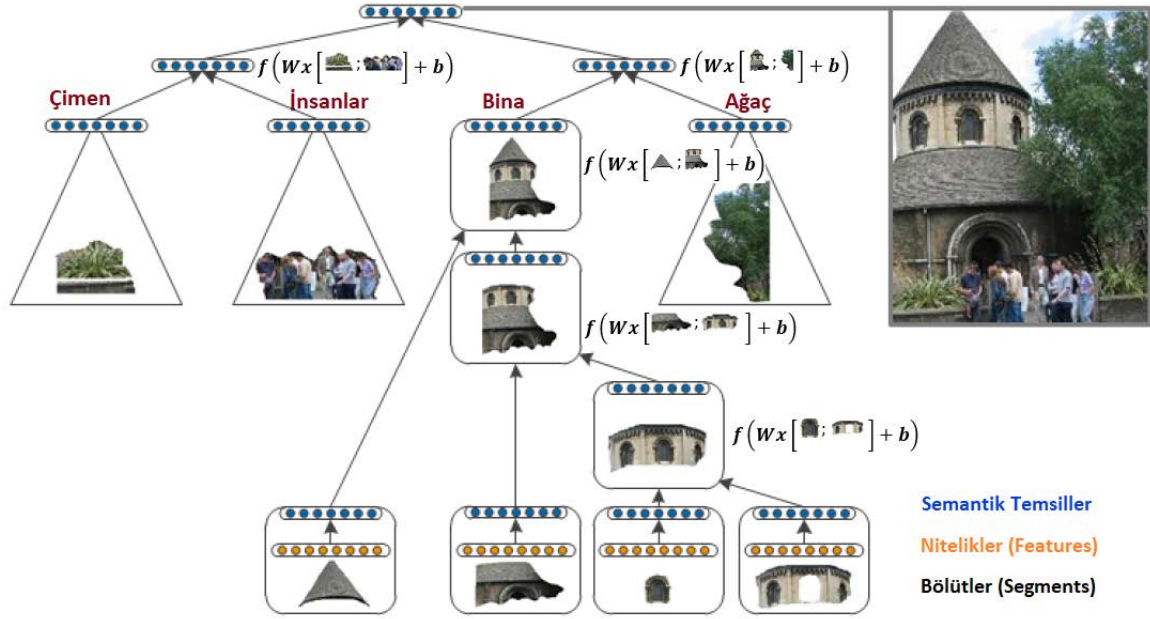
2.1.4. Sonuç

Bu bölümde, tez metninin kendi içerisindeki bütünlüğünü ve anlaşılabilirliğini artırmak için evrimsel sinir ağlarına genel bir bakış sunulmaktadır. ESA mimarisi ve genel kavramları ile ESA'nın eğitilmesine ilişkin özet bilgiler sunulmuştur. ESA, en yaygın kullanılan derin öğrenme tekniklerinden biri olup, nesne tanımadan başka nesne tespiti (örn. [35], [36]), nesne takibi (örn. [37], [38]), metin tespiti ve tanınması (örn. [39], [40]), hareket tanıma (örn. [41], [42]), konuşma tanıma (örn. [43], [44]) ve doğal dil işleme (örn. [45], [46]) gibi birçok problemin çözümünde sıkça kullanılmaktadır. Öte yandan, ESA'nın üstün başarısına rağmen teorik altyapısı, hangi koşulların ve düzen yapıların iyi performans göstereceği ve belirli bir görev için en uygun mimarinin nasıl belirleneceği konularını halen açıklayamamaktadır [16], [47].

2.2. Özyinelemeli Sinir Ağları (ÖSA)

2.2.1. Giriş

Özyinelemeli Sinir Ağları (*Recursive Neural Network*), Tekrarlayan Sinir Ağlarının (*Recurrent Neural Network*) zincir-benzeri diziliminden (*chain-based sequences*) farklı bir çizge veri yapısı ile genelleşmiş bir türünü temsil eden, uzun vadede dağıtılmış temsillerin bağımlılığını parça-bütün hiyerarşisi ile yakalayabilecek ölçüde olabileceği ağaç yapıları üzerinde aşağıdan yukarıya doğru çalışan yapılardır [25], [48], [49]. ÖSA'nın tarihçesi, aslında Pollack [50] ve Hinton'ın [51] çalışmaları ile çok eskiye dayansa da, son yıllarda özellikle Socher ve diğerleri tarafından yapılan çalışmalarla [49], [52]–[56] tekrar yaygınlık kazanmışlardır.



Şekil 2.1. Özyinelemeli sinir ağları, tekrarlayan sinir ağlarının dizilimini ağaç yapısı içerisinde genelleştiren bir yapıya sahiptir. Örnekte görüntünün aşağıdan yukarıya doğru parça-bütün hiyerarşisi görülmektedir. Şekil, [52] çalışmasından uyarlanmıştır.

Şekil 2.1, özyinelemeli bir sinir ağının görüntü bütününü nasıl anlamlandırdığına dair tekrarlı yapıların nasıl çalıştığı hakkında kısa bir örnek vermektedir. En alt düzeyde bölüt nitelikleri ilk önce semantik bölüt temsillerine eşlenir. Daha üst düzeylerde özyinelemeli olarak bölütler bir araya getirilerek nesnelere ve son düzeyde ise nesnelere görüntünün bütününe eşlenir. Her düzeyde semantik temsiller, şekilden görüleceği üzere bir alt düzeydeki nitelik bilgisini kullanarak aşağıdaki gibi elde edilir:

$$a = f(W \times [F_1; F_2] + b) \quad (2.4)$$

Formüldeki W nöronun ağırlık matrisini, b yanlılık değerini (*bias*) ve f ise aktivasyon fonksiyonunu temsil etmektedir.

ÖSA'nın tekrarlayan sinir ağlarına (TSA) göre net bir avantajı, n boyutlu aynı uzunluktaki bir dizi için derinliğin $O(\log n)$ 'e kadar azaltılarak uzun süreli bağımlılıklarla başa çıkmada yardımcı olabilmeleridir [25]. ÖSA'da sonuca bağlanmamış bir sorun, ağacın en iyi şekilde nasıl yapılandırılacağıdır. Bunun için farklı ağaç yapıları kullanılabilir. Örneğin, doğal dil işleme probleminde Socher ve diğerlerinin önerdikleri [54], [55] gibi, bir dil ayrıştırıcısı tarafından cümlenin ayrıştırma ağacının (*parse tree*) yapısına bağlanabilir. Şekil 2.1'de ise görüntü yapıları dengeli bir ikili ağaç halinde ele alınmıştır. Özyinelemeli ağların birçok çeşidi mümkündür (örn. [48], [55] gibi). Tez çalışmaları kapsamında,

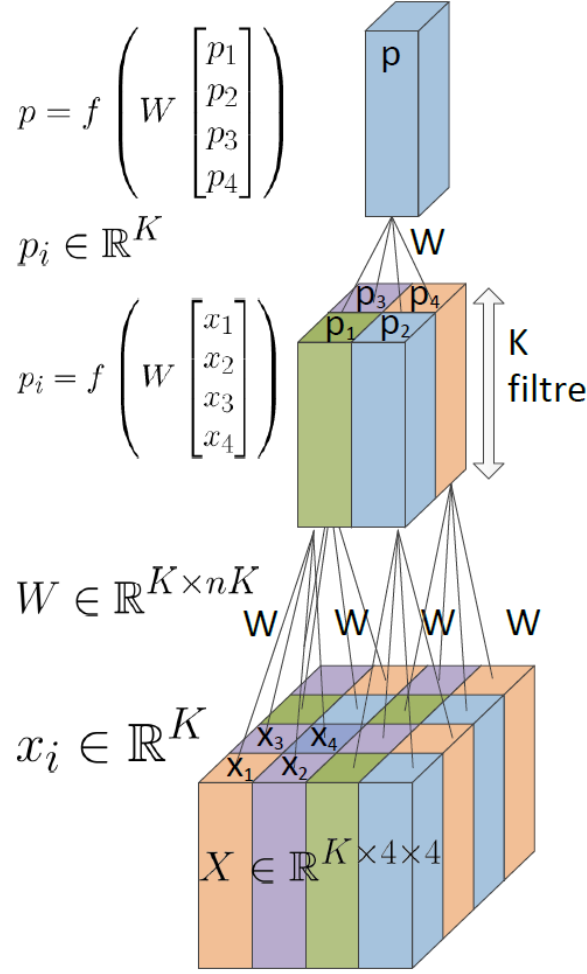
bunlardan çoklu sabit ağaç yapılarına dayanan ÖSA [56] kullanıldığı için, ilerleyen kesimde bu türün mimarisi ele alınacaktır.

2.2.2 ÖSA Mimarisi

Socher ve diğerleri [52], [53] çalışmalarında; giriş verisine bağımlı olan, tek bir tane ve ikili bir yapıdaki ÖSA'nın bir dizi ağırlık değerlerini geriyayılım algoritması ile öğrenen ÖSA modellerinden farklı olarak, [56] çalışmasında giriş verisine bağlı olmadan sabit ve dengeli bir ağaç yapısında çoklu ÖSA modelini önermektedirler. Her bir ÖSA yapısı, giriş verisini yeni bir nitelik vektörüne eşlemekte ve N boyutlu çoklu ÖSA yapılarının oluşturdukları nitelik vektörleri, nihai sonuç vektörünü oluşturmak üzere birleştirilmektedir. Ayrıca ÖSA modellerindeki ağırlık değerleri, geriyayılım algoritması [49] ile öğrenilmek yerine rastgele ilklendirilerek hızlı ve gözetimsiz bir şekilde derin niteliklerinin kodlanması sağlanmaktadır. Jarrett ve diğerleri [57] ile Saxe ve diğerleri [58], çeşitli mimarilerde rastgele ilklendirilen ağırlık değerlerinin, ESA ile nesne tanımda, geriyayılım algoritması ile öğrenilmiş niteliklerle yarışabilecek performans gösterdiklerini sunmaktadırlar. Benzer şekilde, rastgele ilklendirilen çoklu ÖSA mimarisi ile de yüksek kaliteli niteliklerin elde edildiği gösterilmiştir [56]. Ayrıca, geriyayılım algoritmasını ağaç yapısında çalıştırma maliyeti olmaksızın, nitelikler hızlı bir şekilde elde edilerek önemli bir performans avantajı da sağlanmaktadır.

Giriş verisi olarak, 3 boyutlu bir $X \in \mathbb{R}^{K \times r \times r}$ ele alınmaktadır. Her bir kolon vektörü ise K boyutludur. Böylece en alt düzeyde ağacın yaprak düğümlerini, K boyutlu vektörler oluşturmaktadır. Amaç, hiyerarşik bir şekilde alt düzey bloklarını birleştirerek nihai olarak $p \in \mathbb{R}^K$ sonuç vektörünü elde etmektir. Birbirlerine komşu olan blok vektörleri bitleştirilerek ana vektörleri oluştururlar. Bloklar, $K \times b \times b$ boyutludur. Giriş olarak ele alınan X , dengeli ağaca uygun bir yapıda olmalıdır. Örneğin Şekil 2.2'de giriş olarak alınan $X \in \mathbb{R}^{K \times 4 \times 4}$; $X \in \mathbb{R}^{K \times 4 \times 4} \rightarrow X \in \mathbb{R}^{K \times 2 \times 2} \rightarrow X \in \mathbb{R}^{K \times 1 \times 1}$ olmak üzere 3 düzeyli dengeli bir ağaç yapısından oluşmaktadır. Buradaki her bir blok boyutu $K \times 2 \times 2$ 'dir ve her biri K boyutlu olmak üzere (x_1, x_2, x_3, x_4) vektör listesinden oluşmaktadır. Böylece, genel olarak her bir blokta b^2 vektör olmaktadır. Ana vektör ise aşağıdaki gibi hesaplanır:

$$p = f \left(W \begin{bmatrix} x_1 \\ \vdots \\ x_{b^2} \end{bmatrix} \right) \quad (2.5)$$



Şekil 2.2. $K \times 4 \times 4$ ve $K \times 2 \times 2$ boyutlarındaki blokları, giriş çocuk vektörleri olarak ele alan ve her düğümde aynı ağ modelini tekrarlı olarak bir üst vektör olan ana vektörü hesaplamak için kullanan örnek bir özzyinelemeli sinir ağı modeli [56].

Buradaki ağırlık matrisi $W \in \mathbb{R}^{K \times b^2 K}$ formundadır. f ise doğrusalsızlığı ifade eden bir aktivasyon fonksiyonunu ifade etmektedir. Yanlılık değeri ise etkisi olmadığından dolayı kullanılmamaktadır. Formül 2.5, bir ağaç için X matrisindeki tüm bloklara aynı W değerleri ile uygulanmaktadır. Genel olarak, ağaçta yeni bir P_1 matrisini, $(r/b)^2$ sayıda p ana vektörü oluşturur. P_1 'deki vektörler yine Formül 2.5 kullanılarak yeni blokları oluşturmak üzere bitişirilir ve bunlar yeni P_2 matrisini oluşturur. Bu işlem tekrarlı bir yapıda Şekil 2.2'de olduğu gibi tek bir ana vektör kalana kadar devam ettirilir. Örnek olarak elimizde $K = 128$ ve $r = 27$ olmak üzere $X \in \mathbb{R}^{128 \times 27 \times 27}$ giriş verisi olsun. $b = 3$ değeri ile uygun bir şekilde dengeli ağacı oluştururken, her bir blokta $128 \times 3 \times 3$ olur. Yaprak düğümlerdeki her biri 128 boyutlu olan toplamda $27 \times 27 = 729$ sayıdaki vektörler, birleştirilerek $(27/3)^2 = 81$ tane ana vektörünü (p_1 diyelim) oluşturur. Bu düzeydeki $9 \times 9 = 81$ sayısındaki p_1 vektörleri, birleştirilerek $(9/3)^2 = 9$ tane yeni ana vektörü (p_2

diyelim) oluşturur. Daha sonra, $3 \times 3 = 9$ sayısındaki p_2 vektörleri, birleştirilerek $(3/3)^2 = 1$ tane nihai ana p vektörünü oluşturur ve işlem tamamlanır. Sonuç vektörü olan nihai vektör p 'nin boyutu ise $K = 128$ olur.

Tek bir ağaç için yukarıdaki gibi işlemler yapılmaktadır. Çoklu sabit ağaçlı ÖSA mimarisinde ise toplamda N tane ağaç kullanılmaktadır. Her bir ağaç, 3 boyutlu X verisini giriş olarak alıp, işlettikten sonra nihai bir K boyutlu vektör elde eder. Tüm ağaçlardan elde edilen K vektörleri birleştirilerek, en son $N \times K$ boyutlu bir sonuç vektörü elde edilir. Burada sezgisel olarak her bir ağaçta K boyutlu bir filtre veya nitelik elde etme misyonu verilmektedir. Böylece çoklu ağaç yapıları ile N tane ayrı filtre elde edilmiş olur.

2.2.3 Sonuç

Bu bölümde, özellikle son yıllarda artan bir öneme sahip olan özyinelemeli sinir ağları ele alınmıştır. Özyinelemeli modellerin genel yapılarından bahsettikten sonra tez kapsamında kullanılan formu olan, çoklu, sabit ve dengeli ağaç yapısından oluşan ÖSA mimarisi anlatılmaktadır. Saxe ve diğerlerinin [58] çeşitli teorem ve deneysel analizlerle kanıtladığı ESA'daki rastgeleleştirme, ağırlık vektörlerinin rastgele ilklendirildiği ve geriyayılım algoritması kullanılmadan sadece ileri-besleme ile sonuç üreten bu yapılarda da, nesne tanıma probleminde başarılı olduğu söylenebilir.

3. RGB-D NESNE TANIMA İÇİN DERİN NİTELİKLERİN ÖĞRENİLMESİNİN DENEYSEL BİR ANALİZİ

3.1. Giriş

Derin öğrenme yöntemlerinde model ağırlıklarının öğrenildiği eğitim aşaması zaman alan pahalı bir aşamadır. Bu aşamada genellikle büyük bir etiketli veri yığını kullanılarak, gradyan düşüş algoritması ile maliyet fonksiyonu çıktısını minimize etmek için eğitim verisi üzerinden tekrar tekrar geçilir. Ancak model ağırlıklarının öğrenilmediği gözetimsiz öğrenme modelleri de mevcuttur. Derin öğrenme yöntemlerinde genellikle bu modeller bir ya da iki katman içeren sığ modeller olup, özellikle küçük veri kümelerinde başarılı sonuçlar vermiştir (örn. [56]–[60]). Bu tür modellerde ağırlıklar, çoğunlukla rastgele ilklendirilir (örn. [57] ve [58] çalışmalarında olduğu gibi) ya da kümeleme algoritması gibi gözetimsiz bir algoritma ile (örn. [59] ve [60] çalışmalarında olduğu gibi) öğrenilerek kullanılır. Özellikle ImageNet [23] gibi büyük veri kümelerinin yaygınlaşması ve gelişen donanım teknolojisi ile birlikte ucuzlayan eğitim aşaması ile beraber bu tür yöntemlerle son yıllarda daha az karşılaşılmaktadır.

Derin mimariler, katmanların yapılandırması ile bu katmanlarda kullanılan işlevleri içeren genel model özelliklerine ve alıcı alanlar, filtre sayısı ve boyutları gibi “meta-parametreler”e bağlı olarak, farklı karakteristik özelliklere göre çeşitlendirilebilir. Bu bağlamda, son yıllarda birçok sayıda nitelik öğrenme algoritması [56], [60]–[66] önerilmiştir. Tipik olarak, bu yöntemlerin çoğu, hem RGB hem de derinlik görüntüleri için aynı modülerliğe konsantre olmuştur. RGB ve derinlik verilerinin karakteristik özellikleri farklı olduğundan dolayı, bu yaklaşımların uygunlukları şüphelidir. Ayrıca, anılan bu yöntemlerin geriyayılımlı öğrenmeye dayalı olmayan sığ modeller olduğu düşünüldüğünde, giriş verisinin yapısına uygun model parametrelerinin seçimi, uygun nitelik kodlamasını sağlayacaktır. Ancak, hangi veri türüne hangi model parametresinin daha uygun olduğu sorusu, cevap bekleyen bir sorun olarak durmaktadır.

Bu bölümde anlatılan çalışmada, hem RGB hem de derinlik verileri için farklı model parametrelerinin etkileri deneysel olarak araştırılmaktadır. Bu amaçla, literatürde sıkça kullanılan CNN-RNN [56] yöntemi temel alınarak, çeşitli model parametreleri kombinasyonlarını dikkate alan deneyler yapılmaktadır. Bu alanda popüler olan Washington RGB-D Nesne [67] veri kümesini kullanarak, farklı yama çıkartma yaklaşımlarının, doğrultucu birimlerinin, havuzlama yöntemlerinin ve sınıflandırıcıların nihai tanıma üzerindeki etkisi deneysel olarak değerlendirilmektedir. Sonuç olarak,

beklentileri doğrulayan, RGB ve derinlik verileri için farklı model parametrelerinin, aynı model parametrelerini kullanan başlangıç modeline göre daha iyi sonuçlar verdiği gözlemlenmektedir. Dolayısıyla uygun model parametrelerini seçerek, yaklaşımda radikal bir değişiklik yapılmaksızın, RGB-D nesne tanıma başarısının önemli derecede artırılabilirdiği gösterilmektedir.

3.2. İlgili Çalışmalar

Son yıllarda, RGB-D nesne tanıma için birçok nitelik öğrenme yöntemi sunulmuştur. Bu bölüm kapsamında, geriyayılım algoritması kullanılmadan, sığ mimariler kullanarak yapılan öğrenme çalışmaları özetlenmektedir. Bo ve diğerleri [62], HMP olarak adlandırdıkları hiyerarşik eşleme yöntemlerinde, gözetimsiz bir şekilde nitelik öğrenmek için seyrek kodlamayı kullanan bir yaklaşım sunmaktadır. Jhuo ve diğerleri [66], bir nesneye ilişkin gri tonlama ve derinlik görüntüleri arasındaki ilişkiyi belirlemeye çalışan bir ileri-besleme modeli önermektedir. Blum ve diğerleri [60], ilgi noktaları etrafından nitelikleri öğrendikleri evrimsel k -ortalama tanımlayıcısı dedikleri yöntemi önermektedirler. RGB ve derinlik görüntülerini kullanarak ilgi noktaları etrafından sözcük torbası (*bag-of-words*) mantığı içerisinde nitelik öğrenme ele alınmaktadır. Belirlenen ilgi noktalarının komşuluklarındaki nitelik yanıtlarını otomatik olarak öğrenen ve renk, derinlik ipuçlarını bir arada sunan nihai bir RGB-D tanımlayıcı sunulmaktadır. [56] çalışmasında Socher ve diğerleri, RGB ve derinlik görüntülerinden ayrı ayrı nitelik öğrenen ve nihai RGB-D sonucunu bulmak için bu niteliklerin birleştirildiği CNN-RNN olarak adlandırılan evrimsel-özyinelemeli sinir ağlarına dayalı bir yöntem önermektedirler. RGB-D nesne tanıma için zamanında en iyi sonucu veren bu çalışma, daha sonra sıklıkla literatürde çeşitli geliştirmelere (örn. [61], [63], [64] gibi) konu olmuştur. Literatürdeki önemi ve sunulan mimarinin uygunluğundan ötürü bu bölümde anlatılan çalışmada, söz konusu olan bu yöntem temel alınmaktadır. Cheng ve diğerleri [63], CNN-RNN çatısını kullanan yarı-gözetimli bir öğrenme modelini önermektedirler. Daha sonra aynı yazarlar, ESA'nın aynı boyutlarda giriş görüntüsünü alan yapılarından dolayı görüntülerin kırılmaları (*cropping*), biçimsel olarak eğilmeleri (*warping*) yollarıyla, elde edilen yapıların bozulmasını önlemek için bir uzamsal piramit havuzu (*spatial pyramid pooling*) katmanını kullanarak bu çalışmayı genişletmektedirler [64]. Bu çalışmada ayrıca RGB ve derinlik görüntülerine ek olarak gri tonlamalı yoğunluk görüntülerini ve yüzey normallerini de kullanılmaktadırlar. Bai ve diğerleri [61], önerdikleri çalışmalarında filtreleri öğrenmek için yaygın yaklaşım olan rastgele yamaların çıkartılması yerine, veri kümeleri üzerinde tanımladıkları alt

kümelere dayalı bir yaklaşımın kullanılmasını savunmaktadırlar. Guo ve diğerleri [65], sokak görüntülerinden yakalanan apartman numaralarını tanımak için ESA ve Saklı Markov Modelini (SMM) birleştiren entegre bir model sunmaktadırlar. Kayan pencereler (*sliding window*) ile çıkartılan çerçeve dinamiklerini modellemek için SMM ve çerçevelerin görünümünü ele almak için ise ESA'yı kullanırlar. Coates ve diğerleri [59], bu çalışmaya benzer bir analiz çalışmasını RGB alanında sunmaktadırlar. Filtre öğrenmeye ilişkin gözetimsiz birkaç öğrenme yöntemlerini, alıcı alan büyüklüğü, filtre sayısı, adım atlama (*stride*) sayısı gibi “meta-parametre” leri ve öğrenilen niteliklere uygulanan beyazlatmanın (*whitening*) etkisini inceleyen detaylı bir analiz çalışması sunmaktadırlar.

Bu bölümde anlatılan çalışmada, farklı model parametrelerinin RGB ve derinlik görüntülerindeki etkileri ayrı ayrı incelenmektedir. RGB ve derinlik görüntülerinin karakteristikleri farklı olduğu için, her iki veri türü için de aynı modellerin yukarıda anılan geleneksel yaklaşımlarda olduğu gibi kullanılmasının etkili bir çözüm olmadığı sonucuna varılmaktadır.

3.3. Öğrenme Yöntemi

Derin nitelik öğrenme yöntemleri, nitelikleri, düşük düzeydeki ham piksellerden yüksek düzeydeki nesnelere anlamlı parçalarına kadar, hiyerarşik bir şekilde öğrenirler. El tasarımı niteliklerin aksine, derin nitelikler ile genelde daha başarılı sonuçlar elde edilmektedir. Çünkü bu yöntemlerde, ham piksel değerleri arasındaki uzamsal ilişkiyi dikkate alarak, görüntülerin temel düzeyinden öğrenme gerçekleştirilir. Bu bölümde, farklı model parametrelerinin derin nitelikleri öğrenme üzerindeki etkisi, [56] çalışması temel alınarak incelenmektedir. Temel alınan derin öğrenme sistemi, temelde dört birimden oluşmaktadır. İlk aşamasında bir filtre öğrenme modülü bulunmakta, ikinci aşamasını öğrenilen filtreleri ağırlık değerleri olarak kullanan tek bir ESA katmanı oluşturmakta, bu katmanı takip eden aşamada ESA ile öğrenilen yapılar arasındaki üst düzey anlamsal ilişkiyi sağlayacak bir ÖSA katmanı bulunmaktadır. Son olarak ÖSA katmanından elde edilen niteliklerin verildiği, nesne kategorilerini belirleyen bir sınıflandırıcı katmanı mevcuttur. Sistemin genel yapısına ilişkin şematik bir görünüm, Şekil 3.1’de görülmektedir.

3.3.1. Filtre Öğrenme Modülü

Model, etiketsiz giriş görüntülerinden rastgele görüntü yamaları çıkarma işlemi ile başlar. Her bir yama, $w \times w$ boyutlu ve d kanal boyutunu içermektedir (RGB için $d = 3$ ve derinlik görüntüleri için $d = 1$). Daha sonra bu yamalar beyazlatma ve normalizasyon

önişlemlerinden geçirilerek kümeleme algoritması öncesi korelasyonun daha iyi sağlaması amaçlanmıştır. Sonraki adımda, k -ortalama kümeleme algoritması bu görüntü yamalarına uygulanarak K adet merkez öğrenilmektedir. Tüm bu adımlar hem RGB hem de derinlik görüntüleri için uygulanmaktadır. Rastgele görüntü yamalarının çıkartılması, bu yamaların ne kadar ayırt edici ve anlamlı olduğu sorusunu akla getirmektedir. Ayrıca [60] çalışmasında, yazarlar ilgi noktaları etrafında yamalar çıkartarak daha etkin bir sonuca ulaşmaktadırlar. Dolayısıyla, bu modülde ilk adım olarak rastgele yama çıkartma işleminin ne kadar anlamlı olabileceği araştırılmaktadır. Bu amaçla, SIFT [68] ve SURF [69] algoritmaları kullanılarak, tespit edilen ilgi noktaları etrafında yama çıkartılarak daha ayırt edici ve anlamlı filtreler öğrenilebiliyor mu diye sorgulanmaktadır. İkinci olarak, etiketsiz veriden rastgele yamalar çıkartıldığında tüm kategori örneklemlerinden yama çıkartılması garanti edilmemektedir. Bu durum, örneğin sınıf-içi çeşitliliği olan bir kategori için, eğer farklı bir örneklemden yama çıkartılmıyorsa başarıyı azaltabilir. Bu duruma karşı, tüm kategori örneklemlerinden yama çıkartılmasını garantileyecek rastgele yama çıkartımı sağlanmaktadır. Dolayısıyla bu modülde yapılan değişikliklerle, rastgeleliğe karşı sezgisel iyileştirme adımları karşılaştırılmaktadır.

3.3.2. ESA Katmanı

ESA, bir önceki katmandan elde edilen K adet öğrenilmiş filtreyi kullanarak, giriş görüntülerinden temel nitelikler çıkartmak amacıyla kullanılmaktadır. ESA, sabit boyutlu giriş görüntülerinin verilmesini gerektirir. Dolayısıyla giriş görüntüleri eşit $n \times n$ piksellerine yeniden boyutlandırılmaktadır. Böylece $n \times n$ boyutlu ve d kanallı her bir giriş görüntüsü, K adet öğrenilmiş ağırlık değerleri ile evriştirilerek $(n - w + 1) \times (n - w + 1) \times K$ boyutlu evrişimsel çıktılar elde edilmektedir. Ardından, bunu doğrultucu birim ile yerel kontrast normalizasyonu adımları takip eder. Son olarak, ESA'da yaygın bir işlem olarak uygulanan havuzlama adımı ile $(n - w + 1 - p)/s + 1$ boyutlu karesel bölge çıktıları elde edilmektedir. Burada, p havuzlama karesinin boyutunu, s ise atlama adım sayısını ifade etmektedir. Bu işlemler sonucunda, düşük düzeyli değişmez nitelikler öğrenilmiş olmaktadır. Bu adımlar yine hem RGB hem de derinlik görüntüleri için ayrı ayrı uygulanmaktadır.

3.3.2.1. Doğrultucu Birimi

Doğrultucu birim, sayısal giriş değerlerine belirli bir matematiksel işlemi uygulayarak, doğrusal olmayan bir çıktıyı elde etmek amacıyla kullanılır. Temel alınan çatıdaki mutlak değer fonksiyonuna (Denklemler 3.1) ek olarak, derin öğrenme yöntemlerinde sıkça

kullanılan (örn. [22], [70]) *ReLU* (Denklem 3.2) ve *leaky ReLU* (Denklem 3.3) doğrultucu fonksiyonlarının sonuçları araştırılmaktadır.

$$f(w^T x) = |w^T x| = \begin{cases} w^T x, & w^T x \geq 0 \\ -w^T x, & w^T x < 0 \end{cases} \quad (3.1)$$

$$f(w^T x) = \max(0, w^T x) = \begin{cases} w^T x, & w^T x > 0 \\ 0, & w^T x \leq 0 \end{cases} \quad (3.2)$$

$$f(w^T x) = \max(0.01w^T x, w^T x) = \begin{cases} w^T x, & w^T x > 0 \\ 0.01w^T x, & w^T x \leq 0 \end{cases} \quad (3.3)$$

3.3.2.2. Havuzlama

Havuzlama, büyük bir giriş verisinin istatistiklerini, daha düşük bir boyutta özetleyen, derin öğrenme çalışmalarında yaygın kullanılan bir yöntemdir. Giriş verisi düşük boyuta özetlenirken, parametre sayısı önemli derecede azaltılarak hesaplama verimliliği artırılmaktadır. Ayrıca görsel algılama sistemindeki hücrelere benzer bir rolde, küçük çarpıklıklara ve bozulmalara karşı sağlamlık sağlar [57]. Bu çalışmada, iki yaygın kullanılan havuzlama, temel alınan modeldeki “ortalama havuzlama” ile “maksimum havuzlama” yöntemlerinin etkileri karşılaştırmalı olarak araştırılmaktadır. Giriş görüntüsü, p pencere boyutları ile s adım atlama sayısı faktörlerine bağlı olarak, sırasıyla ortalama değerleri ve maksimum değerleri ile uzamsal boyutları boyunca örneklenerek özetlenmektedir.

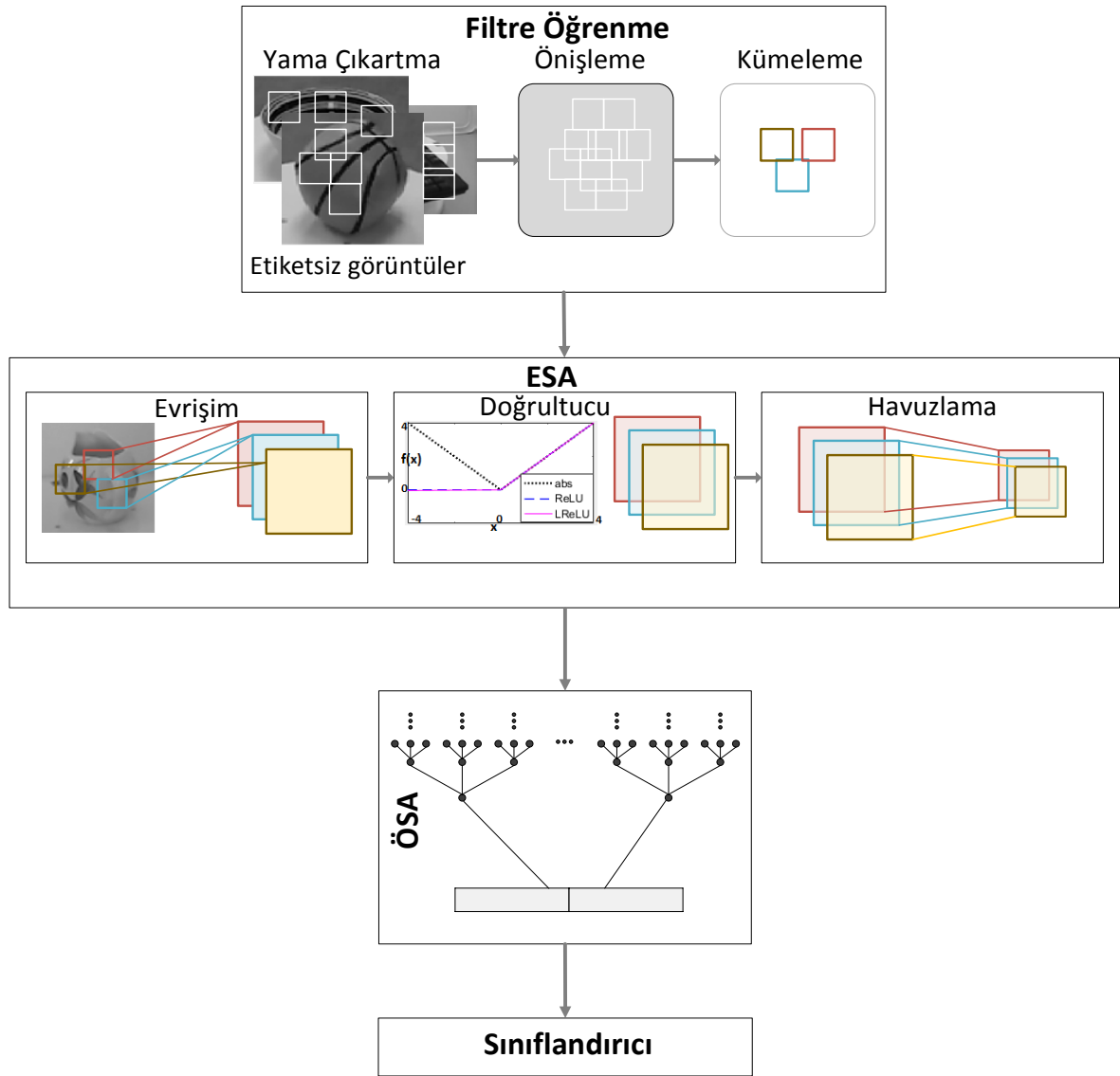
3.3.3. ÖSA Katmanı

ÖSA katmanı, ESA katmanından öğrenilen düşük düzeyli nitelikleri giriş olarak alıp, dengeli sabit ağaçlar vasıtasıyla bu nitelikleri, üst düzey nitelik hiyerarşilerine eşler. Verilen bir 3-boyutlu giriş verisi için, ESA katmanındaki havuzlamada özetlenen $X \in \mathbb{R}^{K \times r \times r}$ çıktısını alarak, yeni bir $y \in \mathbb{R}^K$ gösterimine dönüştürür. Ağaçtaki her bir düzeyde, birbirlerine komşu olan blok vektörleri bir ata vektörde, rastgele ilklendirilmiş aynı ağırlıkları kullanarak birleştirilmektedir. Ata vektörü ise \tanh doğrusal olmayan fonksiyonundan geçirilmektedir. Bu işlem, ağaçtaki tüm düğümler tek bir ata vektörde birleşene kadar, üste doğru özyinelemeli bir şekilde devam ettirilmektedir. Tüm bu adımlar bir tek ağaç yapısı için yapılmaktadır. Bu çalışmada, rastgele ilklendirilmiş ağırlıkları kullanan birden fazla N adet ÖSA kullanılmaktadır. Her bir ÖSA K -boyutlu bir vektör üretmektedir. Dolayısıyla, toplamda $(N \times K)$ -boyutlu bir nihai vektör üretilmektedir. Daha

sonra elde edilen nitelikleri içeren bu nihai gösterimler, tanımlayıcı bir vektör olarak sınıflandırıcıya verilir ve nesne kategorileri bulunur.

3.3.4. Sınıflandırma

Modelin son aşamasını nesne kategorilerini belirleyen sınıflandırma adımı oluşturmaktadır. *Softmax* fonksiyonu, derin öğrenme yöntemlerinde maliyet hesaplamasını yapmak üzere sıkça kullanılmaktadır. Bu çalışmada, *softmax* sınıflandırıcısı kullanılmaktadır. Ayrıca, nihai $(N \times K)$ -boyutlu vektörünü *softmax*'tan başka diğer sınıflandırıcılara vermek de mümkündür. Bu çalışma kapsamında, yaygın kullanılan sınıflandırıcılardan En Yakın Komşu (*Nearest Neighbor* - NN), doğrusal Destek Vektör Makineleri (*linear Support Vector Machines* - SVMs) ve Naive Bayes (NB)' i, *softmax* sınıflandırıcısına ek olarak araştırılmaktadır.

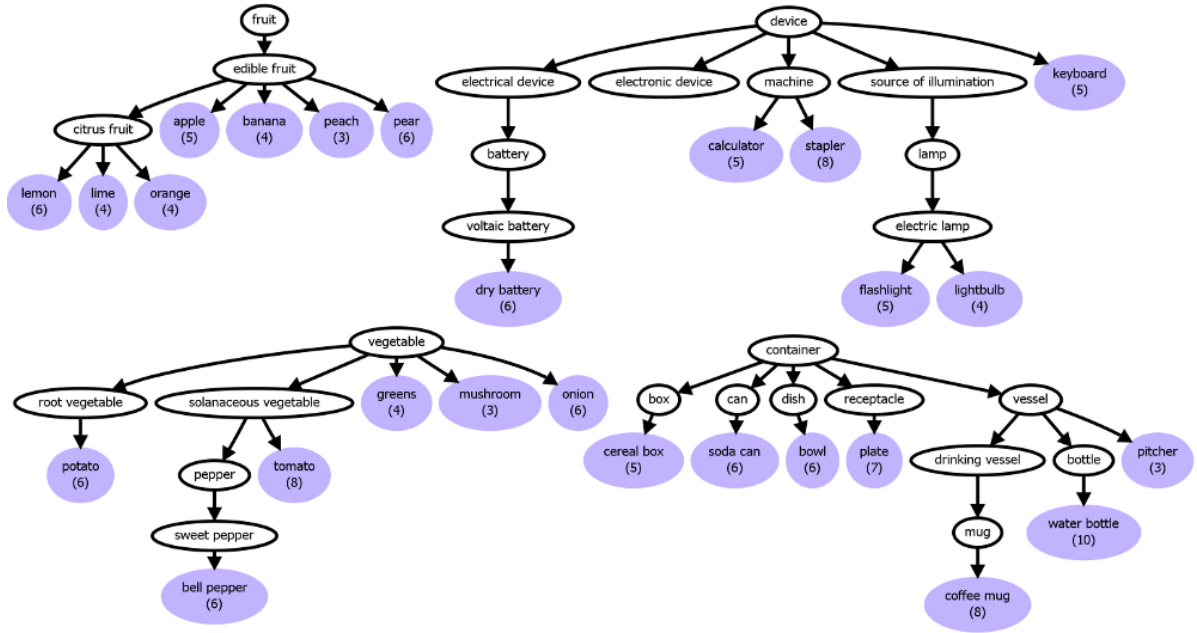


3.4. Deneysel Değerlendirmeler

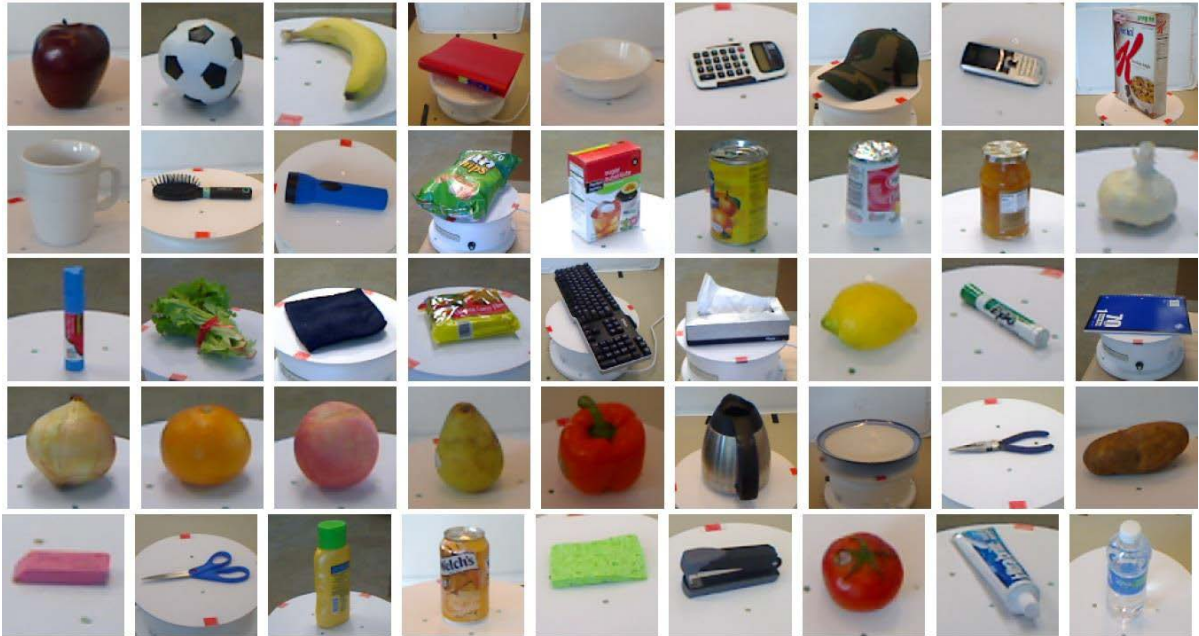
Bu bölümde, çeşitli öğrenme parametrelerinin etkileri yukarıda anlatılan öğrenme modeli çerçevesinde hem RGB hem de derinlik verileri için değerlendirilmektedir. Deneyslerde, RGB-D nesne tanıma için tanımlanmış en eski Kinect veri kümelerinden biri olan ve literatürde yaygın bir şekilde kullanılan popüler Washington RGB-D Nesne veri kümesi [67] kullanılmaktadır. Bu bölümün geri kalan kesimlerinde, ilk önce, bu çalışmada ve tez kapsamında yapılan diğer çalışmalarda kullandığımız Washington RGB-D Nesne veri kümesi hakkında ayrıntılı bilgiler sunulup, bu çalışma kapsamında veri kümesinin nasıl kullanıldığına dair bilgiler verilecektir. Daha sonra filtre öğrenme kapsamında yapılan deneylerin etkilerinden bahsedilecektir. Sonraki kesimlerde, sırasıyla farklı doğrultucu birimlerinin ve havuzlama yöntemlerinin etkileri anlatılacaktır. Son olarak farklı sınıflandırıcı kullanımlarının karşılaştırmalı sonuçları sunulup, deneyler hakkında yapılan değerlendirmelere ilişkin tartışmalar kesimi verilecektir.

3.4.1. Washinton RGB-D Nesne Veri Kümesi ve Kullanımı

RGB-D nesne tanıma çalışmalarında literatürde yaygın kullanılan en geniş ve en kapsamlı veri kümelerinden biri Lai ve diğerleri [67] tarafından tanıtılmış olan veri kümesidir. Bu çalışmada Kinect RGB-D kamerasıyla elde edilen bir nesnenin birden çok, farklı görünümünü kapsayan hiyerarşik bir veri kümesi tanıtılmaktadır. Veri kümesi 51 farklı kategoriden toplam 300 nesnenin görüntülerini içermektedir. Diğer veri kümelerinden farklı olarak bu veri kümesinde, 300 adet günlük sık karşılaşılan nesnelere ait ve farklı görünüm açılarına sahip 250,000 adet görüntü mevcuttur. Veri kümesindeki tüm nesnelere WordNet [71] alt sınıf/üst sınıf ilişkilerine uygun ve ImageNet [23]'te tanımlı yapının bir alt kümesi olacak şekilde kategorilere ayrılmaktadır. 30°, 45° ve 60° derecelik açılarla konumlandırılan kameralar ile her bir nesne sabit hızda dönen bir platform içerisine konularak bir tam dönüş görüntüleri kaydedilmektedir. Böylece her bir nesne için üç farklı açıdan çekilmiş ve her videoda 250 çerçeve olmak üzere toplamda 250,000 adet RGB + Derinlik çerçevesi elde edilmektedir. Şekil 3.2'de ve Şekil 3.3'te elde edilen veri kümesine ait kategorik hiyerarşi ile nesne örnekleri sırasıyla gösterilmektedir.



Şekil 3.2. Dört farklı kategori için RGB-D veri kümesindeki nesnelerin ağaç yapısında gösterimi. Uç düğümlerde yazılan numaralar o kategorideki nesne sayısını (alt kategori örnekleme – *instance*) göstermektedir [67].



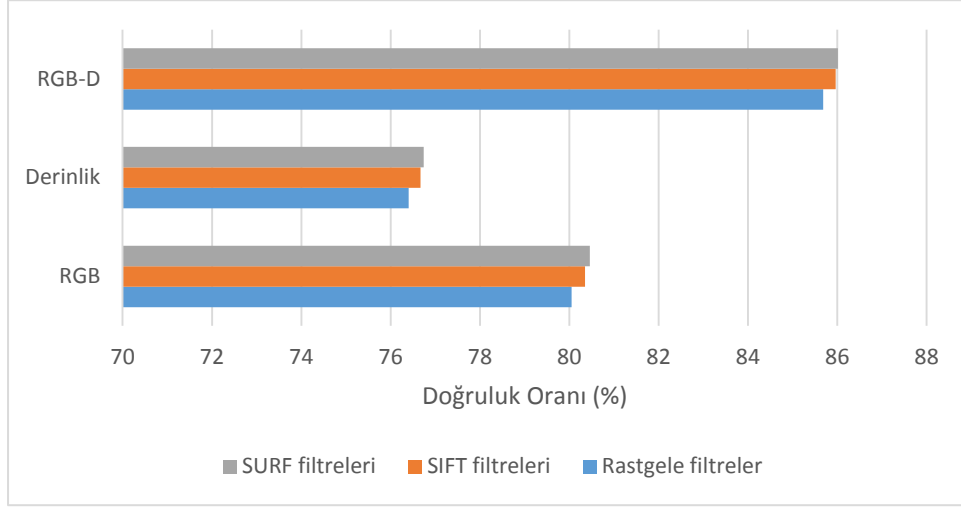
Şekil 3.3. RGB-D veri kümesindeki farklı nesne örnekleri. Burada gösterilen her bir örnek farklı bir kategoriye aittir [67].

Veri kümesi kullanılırken, her bir alt kategori nesnesindeki görüntülerin 5'te biri alınmaktadır. Ayrıca her bir nesne kategorisi için, birer tane alt kategori nesnesi test için ayrılırken, geri kalan nesne alt kategorileri eğitim için kullanılmaktadır. Toplamda 10 farklı eğitim/test kombinasyonu kullanılıp, kategori tanıma başarısı bu kombinasyonların

ortalaması alınarak hesaplanmaktadır. Veri kümesinde, eğitim ve test dışında herhangi bir doğrulama (*validation*) alt kümesi tanımlanmamaktadır. Bu yüzden, farklı model parametreleri araştırılırken, yine [67]'de tanımlanan 10 eğitim/test kombinasyonlarını kullanan ancak alt kategori nesne görüntülerinin 1/5'inin alınması yerine, 1/10'u kullanılarak, model parametrelerinin etkileri doğrulanmaktadır. Böylece, veri kümesiyle sağlanan aynı eğitim/test kombinasyonları kullanılarak ve model parametrelerinin 10 farklı kez çalıştırma sonucu alınarak, daha adil değerlendirme yapılmaktadır. Ayrıca, veri kümesinin 1/10'u alınarak işlemlerin hızlı bir şekilde yapılması sağlanmaktadır. Bununla birlikte, son karşılaştırmada ve farklı model parametrelerinden elde edilen nihai modeldeki değerlendirmede, veri kümesinin 1/5'i alınarak literatür kullanımına uygun karşılaştırma yapılmaktadır. İlk adım olarak, evrişim işlemi için bütün görüntüler 148×148 piksel olarak yeniden boyutlandırılmaktadır ve etiketsiz giriş görüntülerinden toplamda 400,000 yama çıkartılmaktadır. Deneylede, filtre büyüklüğü $w = 9$, filtre sayısı $K = 128$ ve ÖSA sayısı $N = 64$ alınmaktadır. Parametrelerin etkisi değerlendirilirken, başlangıç modelinde adım adım ilgili değişiklikler uygulanarak deneyler yapılmaktadır.

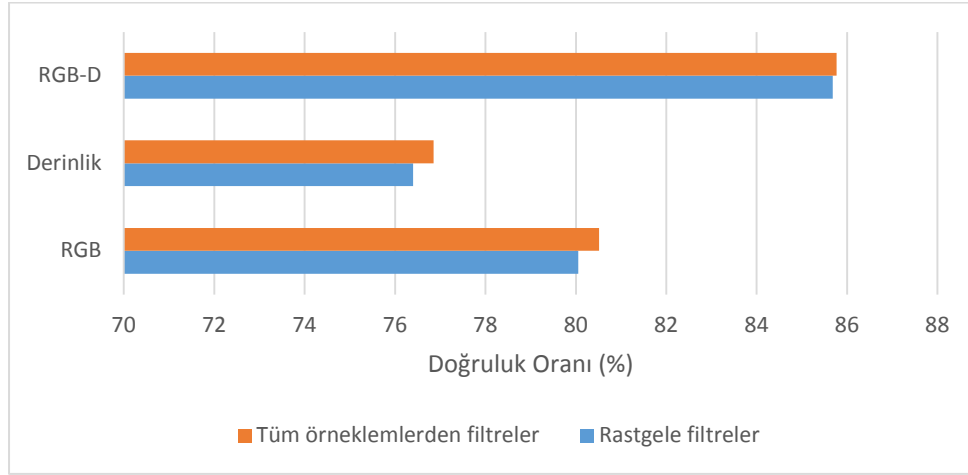
3.4.2. Filtre Öğrenme Yaklaşımlarının Etkileri

Filtre öğrenme modülünde, evrişim ağırlıkları olan filtreler gözetimsiz bir şekilde öğrenilmektedir. Bunun için çıkartılan görüntü yamaları etiketsiz giriş görüntülerinden rastgele bölgelerden çıkartılmaktadırlar. Burada, deneysel olarak araştırılan şey, rastgele bölgelerden görüntü yamaları çıkartmak yerine ilgi noktaları etrafından çıkartılan yamaların daha anlamlı olup olmayacağıdır. Çünkü, SIFT veya SURF ilgi noktaları etrafından çıkartılmış yamalar, görüntüler hakkında önemli ipuçları sağlayan yerel bilgiler sunarlar. Görüntü yamaları, 9×9 (RGB için $9 \times 9 \times 3$ ve derinlik için 9×9) piksel boyutları ile çıkartılmaktadır. Kinect kameradan elde edilen 640×480 tam boyutlu RGB-D görüntülerini kullanmak yerine, veri kümesinde sağlanan döner platformdaki (*turntable*) nesnelere odaklanmış kırılmış görüntüler kullanılmaktadır. Bu yüzden, bazı görüntülerde yeterince ilgi noktaları bulunmayabilir. Bu durumda, rastgele bölgelerden elde edilen görüntü yamaları kullanılmaktadır.



Şekil 3.4. Farklı filtre öğrenme yaklaşımlarının nesne kategorilerini tanımadaki etkisi.

Şekil 3.4'te görüldüğü gibi SIFT yada SURF ilgi noktaları etrafından çıkartılan görüntü yamalarından öğrenilen filtrelerin, rastgele yamalardan öğrenilen filtrelere göre nesne kategorilerini tanımadaki farkı, beklenenin aksine önemsiz denebilecek düzeydedir. Bunun nedeni kırılmış görüntüleri kullanmamız olabileceği gibi, toplamda 400,000 gibi çok fazla sayıda çıkartılan görüntü yamalarından rastgele bile olsa yeterince anlamlı filtre öğrenebileceği sonucu da olabilir.



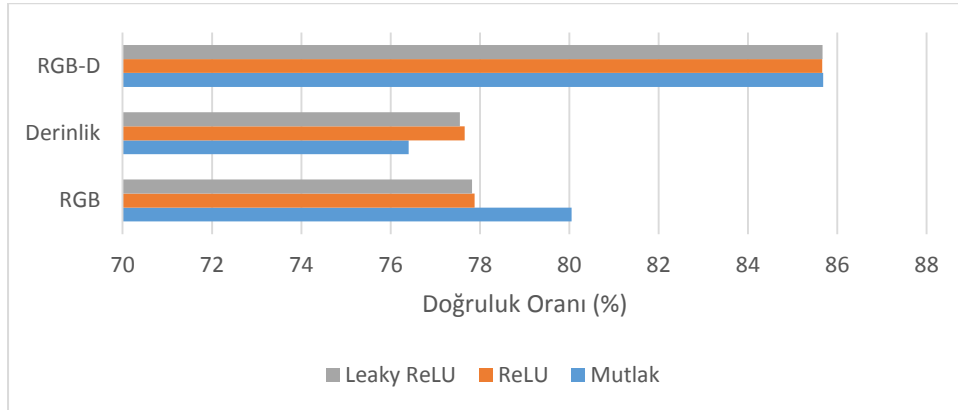
Şekil 3.5. Rastgele giriş görüntülerinden öğrenilen filtreler ile tüm alt kategori nesnelere öğrenilen filtrelerin karşılaştırılması.

Şekil 3.5'te, rastgele giriş görüntülerinden öğrenilen filtrelerin öğrenme başarısındaki etkisinin tüm alt kategori örneklerinden çıkartılan yamalara göre karşılaştırması verilmektedir. Şekil 3.4'e benzer bir sonuca Şekil 3.5'ten de ulaşmak mümkündür. Görüleceği üzere küçük düzeyli bir fark bulunmaktadır. Bu da aslında rastgele çıkartılan

görüntü yamalarından yeterince anlamlı filtrelerin öğrenebileceği sonucuna bizi götürmektedir.

3.4.3. Doğrultucu Birimlerinin Etkileri

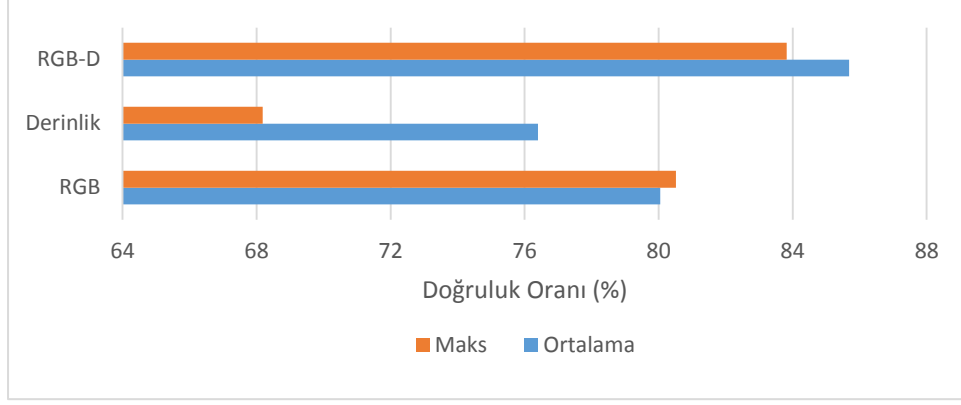
Doğrultucu fonksiyonlarından beklenen, gradyanların kaybolması (*vanishing gradient*) problemi ile karşılaşmadan, yeterli düzeyde doğrusalsızlığı sağlamaları ve hızlı öğrenme için sıfır-merkezli veri dağılımını korumalarıdır. Bu çalışmada kullanılan her üç doğrultucu işlevi de eksenin pozitif tarafında aynı davranırlarken, negatif tarafta ise farklı bir çıktı üretmektedirler. Ancak bu çalışmadaki yöntemde, geri yayılım algoritması ile öğrenme gerçekleştirilmediğinden gradyan problemi ve hızlı yakınsamayı sağlayan sıfır-merkezli veri dağılımının önemi göz ardı edilebilir. Şekil 3.6'da görüleceği üzere, farklı veri türü için farklı doğrultucu fonksiyonları daha iyi sonuç vermektedir. RGB için mutlak değer fonksiyonu daha iyi sonuç verirken, derinlik verileri için ise ReLU ve leaky ReLU fonksiyonları mutlak değer fonksiyonuna göre daha iyi sonuç vermektedir. Aradaki fark şekilden de anlaşılacağı üzere önemli derecededir (RGB için ~2.2% ve derinlik için ~1.3%).



Şekil 3.6. Farklı doğrultucuların etkisi.

3.4.4. Havuzlama Yöntemlerinin Etkileri

Havuzlama için kullanılan pencere boyutları 10×10 ve adım atlama sayısı 5'tir. Böylece havuzlama sonucunda elde edilen çıktılarının boyutları $(148 - 9 + 1 - 10)/5 + 1 = 27$ hesaplaması ile 27×27 olmaktadır.

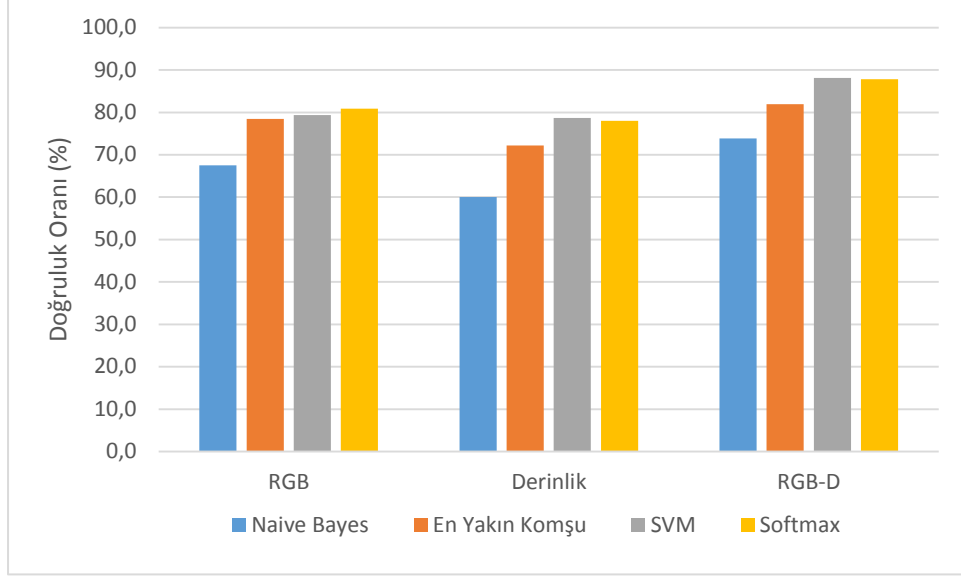


Şekil 3.7. Havuzlama yöntemlerinin etkisi.

Şekil 3.7’de, kullanılan iki yaygın havuzlama yönteminin tanıma başarısının karşılaştırmalı sonucu görülmektedir. Maksimum (Maks) değer havuzlama yöntemi RGB için daha iyi sonuç verirken, ortalama değer havuzlama yöntemi derinlik verileri için büyük bir farkla daha iyi sonuç vermektedir. Bu sonuçlar, veri türüne uygun havuzlama yöntemi seçiminin önemini açık bir şekilde göstermektedir.

3.4.5. Sınıflandırıcı Karşılaştırmaları

Bu kesimde, yukarıdaki deneylere göre RGB ve derinlik için en iyi sonuçları veren parametreleri kullanan model ile üretilen nihai vektörleri, farklı sınıflandırıcı türlerinde sınıflandırarak elde edilen sonuçlar, karşılaştırmalı olarak verilmektedir. RGB için mutlak değer doğrultucu fonksiyonu ve maksimum değer havuzlama kullanılırken, derinlik verileri için ise ReLU doğrultucusu ve ortalama değer havuzlama yöntemi kullanılmaktadır. Görüntü yamaları, SURF ilgi noktaları etrafından çıkartılarak filtreler öğrenilmektedir. Şekil 3.8, sonuçları göstermektedir. Buna göre, SVM ve Softmax diğer sınıflandırıcılara göre daha iyi ve birbirlerine yakın sonuçlar vermektedir. Daha hızlı sonuç alınmasından ötürü son kesimdeki karşılaştırma ayarlarında Softmax sınıflandırıcısı tercih edilmektedir.

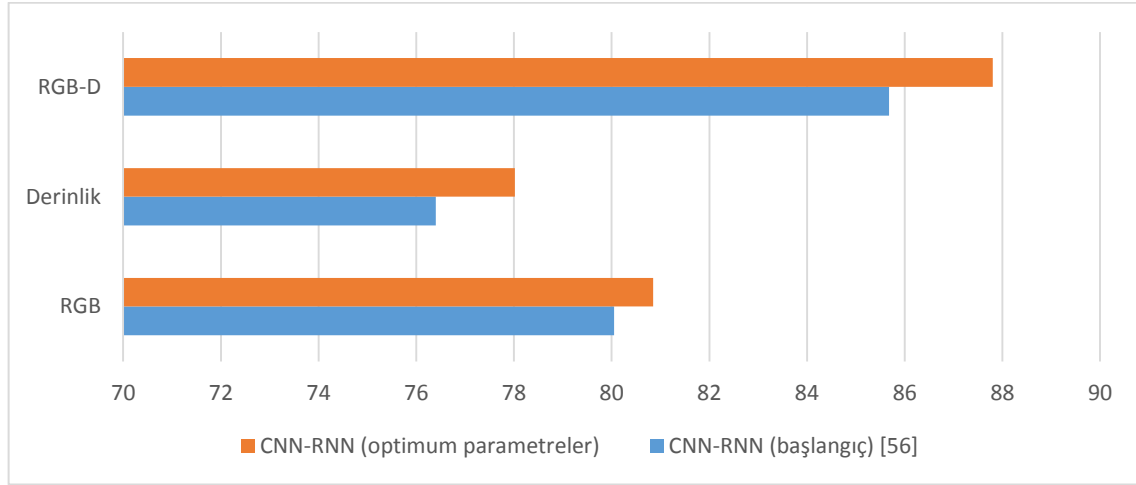


Şekil 3.8. Farklı sınıflandırıcılar arasındaki doğruluk başarısı karşılaştırması.

3.4.6. Tartışmalar

Deneyler, ayrı model parametrelerinin RGB ve derinlik verileri daha iyi sonuçlar verdiğini göstermektedir. Bu beklenen bir sonuçtur. Çünkü RGB ve derinlik görüntüleri, karakteristikleri farklı olan iki ayrı veri türüdür. Ayrıca temel alınan model, geriyayılım algoritması kullanmadan derin nitelikleri kodlayan sığ bir modeldir. Bu yüzden bu tür modellerde uygun parametrelerin önemli bir fark yaratabileceği söylenebilir. Görüntü yamaları çıkartılırken, rastgele yamalar ile SIFT/SURF ilgi noktaları etrafından çıkartılan yamalardan öğrenilen filtreler arasında pek bir fark olmadığı görülmektedir. Filtre öğrenme deneyleri, rastgeleliğin başarılı sonuçlar verdiğini kanıtlamaktadır. Ancak yine de, daha büyük boyutlu giriş görüntüleri için SIFT/SURF gibi ilgi noktaları etrafından yama çıkartmak daha etkili olabilir. Doğrultucu birim için, RGB görüntülerinde mutlak değer fonksiyonu daha iyi sonuçlar üretirken, derinlik görüntülerinde ReLU/leaky ReLU daha iyi sonuç vermektedir. Havuzlama yöntemlerini karşılaştırmak için yapılan deneylerde ise ilginç sonuçlar elde edilmektedir. Maksimum havuzlama RGB’de daha iyi iken, ortalama havuzlamanın derinlik görüntülerinde büyük bir farkla daha iyi olduğu sonucuna ulaşılmaktadır. Son kesimde yapılan sınıflandırıcı karşılaştırmalarında, SVM ve Softmax’ın yakın sonuçlar vererek birbirlerine alternatif olarak kullanılabilceği görülmektedir. Son olarak tüm bu değişiklikleri bir arada kullanan model, temel alınan başlangıç modeli ile karşılaştırıldığında Şekil 3.9’daki gibi bir sonuca ulaşılmaktadır. Bu

sonuçlar, RGB ve derinlik verileri için uygun model parametrelerini seçerek, başarıyı 2%'ye kadar arttırmanın mümkün olduğunu göstermektedir.



Şekil 3.9. Uygun parametrelerinin kullanıldığı modelin başlangıç modeli ile karşılaştırması (%).

Şu ana kadarki uygun model parametrelerini araştıran deneylerde, veri kümesinin yarısı alınmaktadır. Yani her bir alt kategoride alınması gereken 1/5 görüntü yerine 1/10 görüntü alınmaktadır. Ayrıca çıkartılan toplam görüntü yamaları sayısı 400,000 ve ÖSA sayısı 64 olarak ele alınmaktadır.

Çizelge 3.1. Optimum model parametreleri ile ayarlanmış öğrenme modelinin ilgili çalışmalarla olan karşılaştırması (%).

	RGB	Derinlik	RGB-D
EMK-SIFT [67]	74.5 ± 3.1	64.7 ± 2.2	83.8 ± 3.5
KDES [72]	77.7 ± 1.9	78.8 ± 2.7	86.2 ± 2.1
CKM [60]	-	-	86.4 ± 2.3
CNN-RNN [56]	80.8 ± 4.2	78.9 ± 3.8	86.8 ± 3.3
SSL [63]	81.8 ± 1.9	77.7 ± 1.4	87.2 ± 1.1
HMP [62]	82.4 ± 2.1	81.2 ± 2.3	87.5 ± 2.9
Subset-RNN [61]	82.8 ± 3.4	81.8 ± 2.6	88.5 ± 3.1
CNN-SPM-RNN [64]	85.2 ± 1.2	83.6 ± 2.3	90.7 ± 1.1
Bu çalışma	81.8 ± 1.7	79.7 ± 1.9	88.6 ± 1.4

Son olarak bulunan doğru model parametreleri ile ilgili çalışmalar kıyaslaması yapılırken, görüntülerin 1/5'i örneklenerek alınmaktadır. Ayrıca temel alınan modeldeki ayarlara

uygun olarak, görüntü yamaları sayısı 500,000 ve ÖSA sayısı 128 olarak alınmaktadır. Bu deneylerin literatürde ilgili çalışmalarla olan karşılaştırması Çizelge 3.1’de görülmektedir. Çizelgeden de görüleceği üzere uygun model parametrelerini seçerek toplam RGB-D başarısını ~2% artırmak mümkündür. CNN-SPM-RNN [64] çalışmasında olduğu gibi, ayrıca gri tonlamalı görüntüleri ve yüzey normallerini işleterek başarıyı daha da artırmak mümkündür.

3.5. Sonuç

Derin öğrenme yöntemlerinde, nitelik öğrenme aşaması zaman alan pahalı bir süreçtir. Modern ESA mimarileri, onlarca katman ve milyonlarca öğrenilecek parametrelerden oluşmaktadır. Öğrenme süreci geriyayılım algoritması kullanılarak büyük veri kümelerinde tekrar tekrar geçilerek gerçekleştirilmektedir. Bu bölümde, bir ESA ve bir ÖSA katmanından oluşan, geriyayılım algoritması kullanılmadan ileri-beslemeli derin nitelikleri öğrenen bir model üzerinde, deneysel bir analiz çalışması sunulmuştur. RGB ve derinlik verileri için en uygun model parametrelerini bulmak için RGB-D nesne tanıma için tanımlanmış en kapsamlı veri kümesi olan Washington RGB-D veri kümesi kullanılmıştır. RGB ve derinlik verileri için aynı model parametrelerini kullanan diğer ilgili literatür çalışmalarının aksine, deneyler farklı model parametrelerinin RGB ve derinlik görüntüleri için en iyi sonuçları ayrı kullanımlarla ürettiğini göstermektedir. RGB için mutlak değer doğrultucu işlevi ve maksimum havuzlama daha başarılı sonuçları verirken, derinlik görüntüleri için ise ReLU ve ortalama havuzlama en iyi sonuçları üretmektedir. Evrişim filtreleri öğrenilirken, görüntü yamalarının rastgele çıkartılmasının, ilgi noktaları etrafından öğrenilen filtreler kadar başarılı sonuçlar ürettiği gözlemlenmektedir. Doğru model parametreleri seçimi ile sistemin toplam başarısının önemli derecede artırılacağı sonucuna ulaşılmaktadır.

Öte yandan, bu bölümde anlatılan çalışmada, gradyan tabanlı bir öğrenme gerçekleştirilmemektedir. Dolayısıyla nesne tanıma başarısı, genel olarak gradyan tabanlı öğrenme gerçekleştiren yaklaşımlara göre sınırlı olmaktadır. Bunun yanı sıra, bu çalışmada derinlik verileri RGB renk kanallarına ek bir kanalmış gibi ele alınmaktadır. Bu durum, derinlik bilgilerinde saklı olan zengin geometrik bilgilerden tam olarak istifade edilmemesine yol açabilir. Anılan bu dezavantajları aşmak için, bir sonraki bölümde, derinlik verilerini hacimsel temsiller ile ele alan ve 3B ESA ile gradyan tabanlı öğrenme gerçekleştirerek nesne kategorilerini tanıyan çalışmalar yapılmıştır.

4. DERİNLİK VERİLERİNDE 3B ESA KULLANARAK HACİMSEL NESNE TANIMA

4.1. Giriş

3 boyutlu (3B) nesnelere tanıma, otonom robotlardan sürücüsüz araçlara kadar geniş bir uygulama alanına sahiptir. Özellikle robotik alanında, robotun dış dünya ile etkileşimini artırmak için bu tür bir zeka gerekmektedir. Bilgisayarlı görü alanı, görüntülerden gerçek dünyayı anlama çabası olduğundan dolayı, tanıma sistemlerinde 3B bilgisinin kullanılması, bilgisayarlı görü literatüründe ilgi çeken en eski konuların başında gelmektedir [2], [73]–[75]. Gerçek dünya ise üç boyutlu olduğu için, nesnelere 3B gösterimlerini keşfetmek, nesne tanıma alanında önemli bir problem olmuştur. Kinect gibi düşük maliyetli RGB-D algılayıcıların popüleritesi, son yıllarda 3B nesne tanıma hızla ilerleme sağlamıştır. Bu algılayıcılar, güvenilir 3B nesne gösterimleri oluşturmayı sağlayan derinlik ve RGB verilerini bir arada sunarlar. Derinlik verileri aydınlatma, renk, bakış açısı değişikliklerine karşı nispeten daha dayanıklı olduğundan, derinlik bilgisi 3B nesne tanıma çözümlerinde önemli bir rol oynar.

Nesne tanıma alanında önemli gelişmeler kaydedilmesine rağmen, derinlik verilerini kullanarak nesne tanıma hala açık bir araştırma alanıdır. Alandaki mevcut çabaların çoğu (örn. [56], [60]), RGB kanallarına ek olarak derinlik verilerini ekstra bir kanal olarak kullanmaya odaklanmaktadır. Ancak, RGB ve derinlik görüntüleri farklı karakteristiklere sahiptir. RGB verileri, renk ve zengin doku bilgilerine sahipken, derinlik verileri nesnelere 3B yapısal bilgileri ile güçlü bir şekilde karakterize edilmektedir ve aydınlatma koşullarındaki değişikliklere karşı daha sağlam bir yapıdadır. Özellikle robotik alanında, bir robotun karanlıkta etkileşim kabiliyetlerini arttırmak, düşük ışık koşullarıyla ilgili bazı sorunları azaltmak için bu tür bir veriye ihtiyaç vardır [76]. Ayrıca RGB görüntülerinde bakış açısı değişikliklerinden kaynaklanan bazı yanlış sınıflandırma problemleri de derinlik verileri kullanılarak azaltılabilir. Derinlik verilerinden tam olarak yararlanabilmek için, nesne tanıma derinlik verilerine özel yaklaşımlar gereklidir. Bu bağlamda kompakt 3B nesne gösterimleri, etkin performans gösteren algoritmaların tasarlanmasında önemli bir role sahiptir.

Nitelikleri otomatik olarak öğrenme yeteneğine sahip Evrimsel Sinir Ağları (ESA) [22], geleneksel el yapımı niteliklere dayanan nesne tanıma yaklaşımlarını aşarak bilgisayarlı görü alanında son teknoloji bir araştırma sunmaktadır. Daha yakın zamanlarda ise ESA mimarileri 3B veri alanına (örn. [26], [77]) başarıyla genişletilmiştir. Bu yaklaşımlardaki

anahtar etmen, 3B verileri yeterli bir şekilde temsil eden gösterimlerde ele almaktır. Bu amaçla, bu bölümde anlatılan çalışmalarda derinlik görüntülerini kullanarak nesne kategorilerini sınıflandıran yaklaşımlar önerilmektedir. Bu amaçla ilk önce, derinlik verileri üzerinde iki tür hacimsel gösterim tanımlanmaktadır. Daha sonra bu gösterimleri giriş olarak kullanan, nesne kategorilerini hacimsel gösterimlerden tahmin etmek için, 3B geometrik ipuçlarından yararlan 3B ESA modeli önerilmektedir. 3B ESA, evrişimsel sinir ağlarının görüntülerde devrim niteliğindeki başarısını takiben, video analizi [42] için ilk önerilmiştir. Videolarda hareket bilgisini yakalamak için, birden çok bitişik kareye 3B ESA uygulanmıştır. Bu nedenle, zaman, bu tür video tabanlı 3B ESA uygulamalarında üçüncü boyut olarak rol almaktadır. Daha sonra 3B ESA, ShapeNet [26] ve VoxNet [77] çalışmalarında, hacimsel nesne tanımaya başarılı bir şekilde genişletilmiştir. Bu yöntemlerdeki başarıyı takiben ve 3B mesh modellerini içeren sentetik ModelNet [26] veri kümesinin tanıtımı ile beraber, bazı diğer 3B ESA tabanlı yöntemler (örn. [78], [79] gibi) önerilmiştir. Bu yöntemler, nesnelerin tam görünümünü kapsayan sentetik CAD veri kümesi olan ModelNet üzerinde tanıma yapmaktadır. Öte yandan bu bölümde anlatılan çalışmalarda, ev içi gerçek nesnelere üzerinde, Kinect algılayıcısı tarafından elde edilmiş derinlik görüntülerini kullanan 3B ESA modelleri kullanılmaktadır. Microsoft Kinect'in sağladığı derinlik görüntüleri, nesnelerin 3B yapısı hakkında bakış açısı tabanlı eksik bilgiler sunmaktadır. Yani nesnelerin arka tarafları bilinmemektedir. Bu yüzden nesnelerin farklı açılardan çekilmiş görüntülerini bir araya getiren ikinci bir öğrenme yaklaşımı da önerilmektedir. Bu yaklaşım, bakış açısı ile ilgili belirsizlikleri aşmak için çevresiyle dinamik olarak etkileşime giren bir robot için özellikle önemlidir. Robot yeni bakış açılara ilerledikçe, farklı perspektiflerden kazanılan ipuçlarını bir araya getirerek tanıma başarısını büyük ölçüde artırabilir. Ayrıca giriş görüntülerini dönme matrisleri ile çoğaltarak benzer yaklaşımla ele alan deneyler de sunulmaktadır. Sadece derinlik görüntülerini kullanan bu yaklaşımların yanı sıra RGB verilerini de farklı yaklaşımlarla hacimsel tanımda ele alan deneysel çalışmalar da yapılmaktadır. Bu kapsamda, (i) nesnelerin gri tonlamalı hacimsel gösterimlerini ele alan, (ii) 24-bitlik RGB renk bilgisini 8-bitte kodlayan, (iii) RGB renklerini hiperkübik temsillerde ele alan, (iv) RGB renk kanallarını çoklu-dönme 3B ESA ile ele alan, 4 farklı yaklaşım ve deneysel analizleri sunulmaktadır. Bu bölümde yapılan çalışmalar aşağıdaki gibi özetlenebilir:

- Derinlik verilerini hacimsel olarak ifade eden iki tür hacimsel grid temsilleri önerilmektedir.

- Deneysel olarak çeşitli 3B ESA modelleri araştırılıp, aşırıuyumlamayı azaltan ve hacimsel grid temsillerini daha iyi genelleştiren bir 3B ESA modeli önerilmektedir.
- Tek bir derinlik görüntüsünün kullanımının yanı sıra, nesnelerin farklı açılardan çekilmiş görüntülerini bir araya getiren bir yaklaşım (çoklu-dönüştü) sunulmaktadır. Bu şekilde, ağda çoklu dönme nesne görüntülerinden toplanan bilgiler performansı önemli derecede artırabilir. Yaygın kullanılan iki veri kümesi, bu amaca uygun bir kullanımda ele alınmaktadır. Bu amaçla, Washington RGB-D veri kümesinin bir alt kümesi olan yeni bir veri kümesi oluşturulmaktadır.
- Giriş görüntülerini dönme matrisleri ile çoğaltan ve çoklu-dönüştü 3B ESA yaklaşımına benzer ele alan bir yaklaşım (çoklayan-dönüştü) önerilmektedir.
- Sadece derinlik bilgisini kullanan hacimsel gösterimlerin yanı sıra, RGB renk bilgilerini de hacimsel gösterimler içerisinde ele alan yaklaşımlar önerilmektedir.
- RGB-D nesne tanıma probleminde, yaygın kullanılan iki veri kümesinde nesnelere hacimsel gösterimlerle ele alıp 3B ESA ile tanıma gerçekleştiren, literatürdeki ilk çalışmalar sunulmaktadır. Önerilen yöntemin etkinliğini göstermek için farklı senaryolar üzerinde çeşitli deneyler yapılmaktadır. Washington RGB-D [67] ve 2D3D Nesne [80] veri kümelerindeki deneysel sonuçlar, tek-dönüştü yaklaşımın ilgili diğer yöntemler ile rekabetçi sonuçlar ürettiğini gösterirken, çoklu-dönüştü yaklaşımın dönme değişmezliğini sağlayarak tanıma doğruluğunu daha da artırdığını göstermektedir.

4.2. İlgili Çalışmalar

RGB-D algılayıcılarının kullanımlarıyla beraber, derinlik görüntülerini kullanarak nesne tanıma gerçekleştiren yaklaşımlar yaygınlaşmıştır. Bu yaklaşımları, uygulanan teknikleri bakımından üç kategoride ele almak mümkündür: el yapımı nitelik tabanlı derin öğrenme öncesi çalışmalar [67], [72], [80]–[82], 2.5B ESA tabanlı yöntemler [56], [60], [61], [63], [64], [83] ve 3B ESA tabanlı yöntemler [13], [14], [26], [77], [78].

4.2.1. El Yapımı Nitelik Tabanlı Yöntemler

El yapımı nitelik tabanlı yöntemler, geleneksel nitelik çıkartma ve çıkartılan nitelikleri özlu bir temsilde toplama işlemine dayanır. Bu yöntemlerde, çoğunlukla RGB alanında tanımlanmış SIFT [68], SURF [69] ve HOG [84] gibi yerel nitelik tanımlayıcıları derinlik alanına uygulanmaktadır.

Lai ve diğerleri [67], RGB-D nesne tanıma alanında temel kabul görmüş büyük ölçekli bir veri kümesini sunarlarken, bir dizi standart renk ile şekil niteliklerini çıkartarak

sınıflandırma işlemini gerçekleştirmektedirler. Bo ve diğerlerinin derinlik çekirdek tanımlayıcıları [72], derinlik haritasında ve 3B nokta bulutunda birbirini tamamlayan farklı nitelikleri yakalayan boy, şekil ve kenar gibi önemli tanıma ipuçlarını tümleşik bir şekilde sunmaktadır. Hiyerarşik çekirdek tanımlayıcıları [81], derinlik çekirdek tanımlayıcılarını [72], piksel düzeyden nesne düzeyine kadar katmanlı bir düzeyde niteliklerin hiyerarşik olarak toplandığı bir yolla ele almak suretiyle genişletmektedir. Ayrıca kullanılan standart niteliklerin yanı sıra alternatif olarak, [67] ve [72] çalışmalarında yazarlar, 3B yerel şekil tanımlayıcıları olarak spin görüntülerini [85] kullanılmaktadırlar. Bunlardan başka, derinlik görüntülerinde 3B geometrik karakteristikleri yakalamak için derinlik görüntülerine özgün tanımlanan HONV [82] çalışması, standart RGB niteliklerini derinlik alanına uygulayan çalışmalara göre daha başarılı sonuçlar vermektedir. Sonuç olarak, bahsedilen tüm bu çalışmaların odağında, özenli bir el emeği nitelik çıkartımı söz konusudur. ESA gibi son yıllarda geliştirilen derin öğrenme teknikleri, bu tarz el emeği nitelik çıkartımı ihtiyacını ortadan kaldırmıştır.

4.2.2. 2.5B ESA Tabanlı Yöntemler

2.5B ESA tabanlı yöntemlerde, derinlik bilgisi RGB kanallarına ek bir kanal olarak ele alınmaktadır. Bu kapsamda, Blum ve diğerleri [60] tarafından, SURF ilgi noktaları etrafındaki yamalardan gözetimsiz bir şekilde nitelik öğrenen evrışimsel k -ortalama (CKM) tanımlayıcısı sunulmaktadır. Benzer şekilde, CNN-RNN [56] çalışması, evrışimsel sinir ağları ve özyinelemeli sinir ağlarına dayalı bir yöntem sunmaktadır. Hem CKM hem de CNN-RNN nitelik öğrenme yöntemleri, rastgele seçilen görüntülerden çıkartılan yamalardan filtre öğrenmeye odaklanmaktadır. Bu yaklaşımı geliştirmek için, [61] çalışmasında görüntüler şekil ve renk bilgilerine göre alt kümelerle ayrıştırılarak altkümelerden seçilen görüntülerden filtre öğrenilmektedir. Bu yaklaşımla, rastgele öğrenilen filtrelere göre daha iyi bir sonuç elde edilse de, altkümelerin otomatik olarak elde edilmediği not edilmektedir. Cheng ve diğerleri [63], CNN-RNN ile birlikte sunulan bir eş-egitim (*co-training*) algoritmasıyla etiketsiz verileri kullanan yarı gözetimli bir yöntem önermektedirler. Aynı yazarlar, daha sonra bu yöntemi farklı ölçekli görüntüleri ele alacak şekilde SPM (*Spatial Pyramid Pooling*) [86] ile geliştirerek, [64] çalışmasında sunmaktadırlar. Ayrıca RGB ve derinlik görüntülerine ek olarak gri tonlamalı görüntüler ve yüzey normalleri kullanılmaktadır. Zaki ve diğerleri tarafından önerilen Hypercube [83] yöntemi, derinlik verilerinin kullanımında büyük ölçekli RGB veri kümelerinden yararlanmak amacıyla, derinlik haritalarını ve nokta bulutu verilerini RGB gibi üç kanallı

bir yapıda kodlamak için bir yöntem sunar. Böylece ImageNet gibi büyük ölçekli RGB veri kümesinde eğitilmiş bir modeli, derinlik alanında da kullanabilmektedir. Ayrıca kullanılan öneğitimli ESA modelinin sadece son katman aktivasyonları değil, tüm katmanlarından elde edilen aktivasyonları birleştirilerek sınıflandırılma gerçekleştirilmektedir. Dolayısıyla, erken katmanlardan gelen uzamsal olarak birbirlerini destekleyen nitelikleri son katmanlardaki anlamsal bilgilendirici niteliklerle birleştirilerek önemli bir sınıflandırma başarısı elde edilmektedir.

4.2.3. 3B ESA Tabanlı Yöntemler

2.5B ESA tabanlı yöntemler, geleneksel el emeği nitelik tabanlı yöntemlere göre başarı derecesini önemli derecede artırmaktadırlar. Ancak bu yöntemler, derinlik verilerindeki hacimsel bilgileri tam olarak açığa çıkartamamaktadırlar. Nesne tanıma problemi için geometrik bilgileri 3B olarak ele alıp hacimsel olarak tanıma gerçekleştiren ilk çalışma Wu ve diğerlerinin önerdikleri ShapeNet [26] çalışmasında sunulmaktadır. Yazarlar, bu çalışmada hem 3B CAD modellerini içeren ve ModelNet olarak adlandırılan bir veri kümesini tanıtmakta, hem de bu modellerde tanıma gerçekleştiren bir 3B ESA mimarisi önermektedirler. Daha sonra, VoxNet [77], daha az parametre içeren daha küçük bir 3B ESA modeli ile ShapeNet doğruluğunu önemli derecede artırmaktadır. ShapeNet ve VoxNet başarılarını takiben, bu bölümde anlatılan [13] ve [14] çalışmalarında, iki yaygın kullanılan Kinect RGB-D nesne veri kümeleri olan Washington RGB-D [67] ve 2D3D Nesne [80] veri kümelerinde, ilk hacimsel 3B ESA tabanlı yöntemler sunulmaktadır. Bu metinde, bu çalışmalar genişletilerek anlatılmaktadır.

4.3. Önerilen Yöntem

Önerilen yöntemde, Kinect'ten elde edilen ham derinlik görüntüleri kullanılmaktadır. Bu derinlik görüntüleri, gürültülü olup, yansılardan, nesne yüzeylerinin saydamlağından vs. kaynaklı eksik değerler içermektedirler [10]. Görüntülerdeki bu eksik değerler, aynı zamanda delik (*hole*) olarak da bilinmektedir. Bu nedenle, derinlik haritalarındaki eksik değerleri doldurmak için, hedef değer 5×5 'lik piksel komşuluğuyla iteratif bir işlemi ve nokta bulutundaki gürültülerden kurtulmak için de [87] algoritması, önışlem adımları olarak uygulanmaktadır. Önerilen yaklaşım, iki temel adımdan oluşmaktadır. İlk olarak, derinlik görüntülerinden hacimsel temsiller tanımlanmaktadır. Daha sonra, bu hacimsel temsilleri giriş olarak ele alan bir 3B ESA ile tanıma işlemi gerçekleştirilmektedir.

4.3.1. Hacimsel Temsiller

RGB ve derinlik görüntülerinin algısal yapıları farklı olduğu için, derinlik verileri, RGB kanalları ile birlikte ek bir kanal olarak ele alındıklarında (örn. [56], [60]), derinlik verilerinde saklı olan nesnelerin geometrik silüetleri tam olarak ortaya çıkmayabilir. Ayrıca, hacimsel temsiller ESA mimarilerinde avantajlara sahiptir. Nokta bulutundan ve mesh modellerden farklı olarak, hacimsel temsillerin ESA'lara uygulanabilirlikleri ve uzamsal komşuluk ilişkilerini doğrudan tutmayan basit yapıları, bu iki veri türüne göre kullanımlarında anahtar faktörlerdir. Bu amaçla, derinlik verilerine dayalı iki basit ve etkili hacimsel temsil önerilmektedir. Her bir hücresinin bir voksel (hacimsel piksel-*volumetric pixel*) olarak ifade edildiği bu temsiller, nokta bulutu verilerinin 3B matris uzayına yansıtılmalarıyla elde edilmektedirler.

4.3.1.1. İkili Grid (*Binary Grid*)

İkili grid modelinde, bir yüzey noktasının temsil edildiği her bir vokselin ikili bir durumu vardır. Voksel değeri 1 ise temsil edilen alan için bir yüzey değeri var demek olurken, 0 ise yüzey değerinin yokluğunu ifade etmektedir. Verilen bir $P = \{p_1, p_2, \dots, p_m\}$ nokta bulutu için, her bir nokta $p_n = \{x_n, y_n, z_n\}$ ile temsil edilir. Buradaki x_n, y_n, z_n değerleri sırasıyla x, y, z eksenlerindeki 3B koordinat değerlerini temsil etmektedir. m nokta bulutundaki noktaların sayısını ifade etmektedir. Ardından, nokta bulutundan hacimsel gride dönüşüm aşağıdaki gibi gerçekleştirilir:

$$\begin{aligned}x'_n &= \left(\frac{x_n - x_{min}}{(x_{max} - x_{min}) + \varepsilon} \right) (t_{max} - t_{min}) + t_{min} \\y'_n &= \left(\frac{y_n - y_{min}}{(y_{max} - y_{min}) + \varepsilon} \right) (t_{max} - t_{min}) + t_{min} \\z'_n &= \left(\frac{z_n - z_{min}}{(z_{max} - z_{min}) + \varepsilon} \right) (t_{max} - t_{min}) + t_{min}\end{aligned} \tag{4.1}$$

Burada, x'_n, y'_n, z'_n ; p_n 'ye karşılık gelen gride öngörülen voksel pozisyonunu temsil eder. Nokta bulutu verilerinin x, y, z eksenlerindeki maksimum ve minimum değerleri, sırasıyla (x_{max}, x_{min}) , (y_{max}, y_{min}) ve (z_{max}, z_{min}) ile ifade edilmektedirler. Kullanılan grid modeli $30 \times 30 \times 30$ 'luk olduğundan dolayı, t_{max} ve t_{min} olarak ifade edilen hacimsel

gride yansıtma değerleri sırasıyla 30 ve 1 olarak alınmaktadır. Paydadaki sabit $\varepsilon \approx 0$, nokta bulutu eksenlerindeki maksimum ve minimum değerleri eşit olduklarındaki olası sıfıra bölünme problemini önlemek için kullanılmaktadır. x'_n, y'_n, z'_n değerleri, grideki (t_{min}, t_{max}) aralığındaki kesikli değerlerine uygun olarak en yakın tamsayı değerlerine yuvarlanmaktadır. Oluşturulacak grid modelinde, noktaların birbirlerine göre olan pozisyonlarının korunması için x'_n, y'_n, z'_n ayrı ayrı olarak kendi içlerinde hesaplanmaktadır. v_{ijk} 'nin i, j, k koordinatında temsil edilen bir grid vokseli olduğunu varsayalım. Bu durumda, v_{ijk} değeri aşağıdaki gibi hesaplanır.

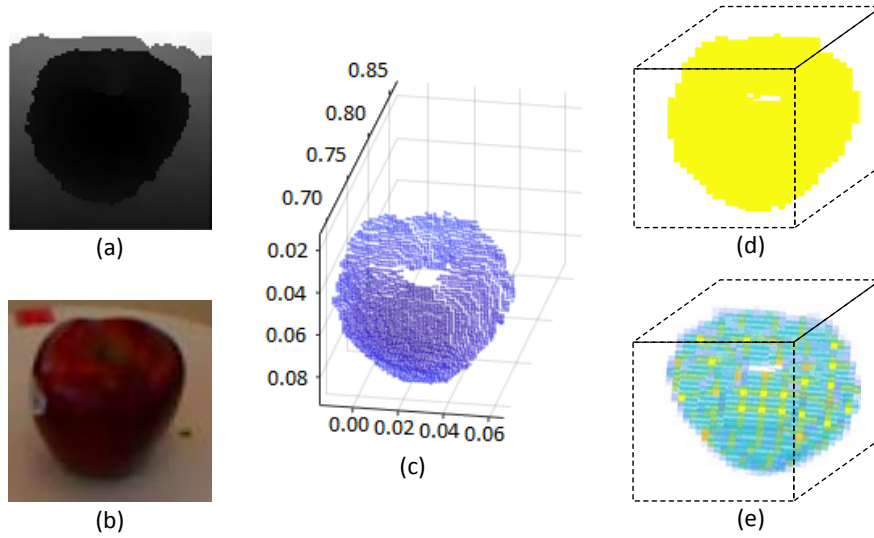
$$\tau_{ijk} = \begin{cases} 1, & \text{eğer öyle bir } p_n \text{ var ki } i = x'_n, j = y'_n, k = z'_n \\ 0, & \text{aksi durum} \end{cases} \quad (4.2)$$

4.3.1.2. Yoğunluk Gridi (*Intensity Grid*)

İkili grid, bir vokselde ilgili bir yüzey noktasının olup olmadığını gösteren basit bir modeldir. Ancak, birçok nokta bulutu değeri aynı voksele karşılık gelebilir. İkili grid yapısında bu sayıdan bağımsız, yüzey noktasının olup olmadığı bilgisi tutulmaktadır. Bu nedenle önerilen hacimsel yoğunluk gridi, bir vokselde bir noktanın varlığını/yokluğunu tutmak yerine, o vokselin kaç noktayı temsil ettiği bilgisini tutmaktadır. Her vokselde, bu voksele yansıtılan noktaların sayısına göre bir yoğunluk değeri hesaplanır. Böylece, bu model ile nesne şekilleri hakkında daha ayrıntılı bilgi elde edilebilir. Bu amaçla, her vokselin değeri sıfır ile başlatılır ve bir voksel içine düşen nokta sayısı (4.3) 'te olduğu gibi güncellenir.

$$\tau_{ijk} = \begin{cases} \tau_{ijk} + 1, & \text{eğer öyle bir } p_n \text{ var ki } i = x'_n, j = y'_n, k = z'_n \\ \tau_{ijk}, & \text{aksi durum} \end{cases} \quad (4.3)$$

Şekil 4.1'de hacimsel temsillerin yapıları gösterilmektedir. İkili grideki değerler sarı renkle temsil edilmektedir. Yoğunluk gridi için yoğunluk değerleri arttıkça renkler koyulaşmaktadır.



Şekil 4.1. Elma nesne kategorisine ait örnek hacimsel gösterimler. (a) Ham derinlik görüntüsü. (b) Derinlik görüntüsüne karşılık gelen RGB görüntüsü. Önerilen yöntemde sadece derinlik görüntüleri kullanılmaktadır. Buradaki RGB görüntüsü görsel amaçlı kullanılmaktadır. (c) İlgili nokta bulutu görünümü. (d) Hacimsel ikili grid. (e) Hacimsel yoğunluk gridi.

4.3.2. 3B ESA

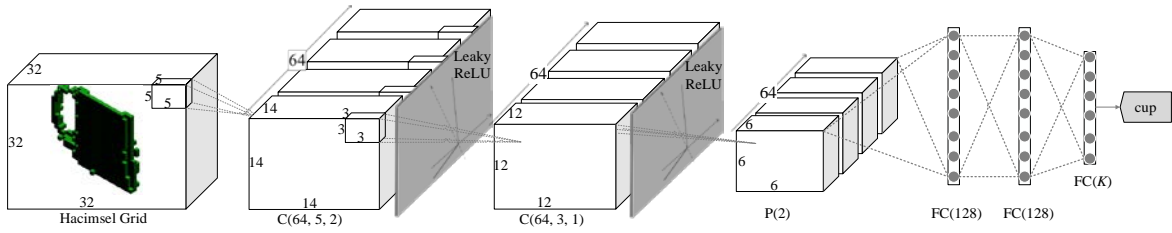
Hacimsel grid temsillerinin oluşturulmasından sonra, bunları giriş verisi olarak ele alan 3B ESA kullanılmaktadır. 3B ESA'nın, 2B ESA veya derinlik bilgisinin ek kanal olarak kullanıldığı 2.5B ESA'lara göre farkı, evrişim işlemi yaparken filtrelerin giriş verilerinde sadece en (*width*) ve boyda (*height*) değil, aynı zamanda derinlik (*depth*) uzamı boyunca da kayırılmasıdır. Böylece tek bir filtrenin giriş verisi ile evriştirilmesinden ortaya çıkacak olan çıkış aktivasyonu, diğer ESA türlerinden farklı olarak yine hacimsel bir 3B aktivasyonu olacaktır. Bu farklılık dışında, 3B ESA bileşenleri yaygın diğer ESA türlerindeki gibidir. Bu çalışma kapsamında kullanılan 3B evrişimsel sinir ağlarının temel bileşenlerini aşağıdaki bileşenler oluşturmaktadır;

- i. Evrişim katmanı $C(k, w, s)$, $w \times w \times w$ boyutlarındaki hacimsel giriş verisine k adet evrişim filtresini, s voksel kayırma adımı ile uygulamaktadır.
- ii. Havuzlama katmanı $P(w)$, daha geniş olan hacimsel giriş verisinin istatistiklerini, verilerin tüm eksenlerindeki uzamsal boyutlarında w katsayı faktörü ile, daha küçük bir çıkış temsili içerisine maksimum değerleri alınarak özetlemektedir.
- iii. Tam-bağlantılı katman $FC(K)$, K adet çıkış nöronu içermekte ve her bir nöronu bir önceki katmandaki tüm aktivasyonlara tam bağlantılı olmaktadır.

Bunların yanı sıra, aşırıuyumlamayı önlemek için p nöron düşürme olasılık faktörü ile $D(p)$ seyreltme (*dropout*) [32] işlemi ile evrişim işlemlerinden sonra doğrusalsızlığı sağlamak amacıyla *leaky ReLU* [70] aktivasyonu kullanılmaktadır. Hacimsel temsillerden nesne kategorilerini tanımak amacıyla, ilk önce VoxNet [77] mimarisi ile öğrenme gerçekleştirilmektedir. Kullanılan 3B ESA modeli, $C(32, 5, 2) - C(32, 3, 1) - P(2) - FC(128) - FC(K)$ temel katmanlarından oluşan, basit ama etkili bir mimari olup, Lasagne [88] kullanılarak geliştirilmiştir. Daha sonra çeşitli alternatif modeller içerisinde deneysel olarak en iyi sonuç veren $C(64, 5, 2) - C(64, 3, 1) - P(2) - FC(128) - FC(128) - FC(K)$ modeli kullanılarak çalışmalar gerçekleştirilmiştir. Anılan bu iki model dışında deneysel olarak araştırılan modellerden bazıları aşağıdaki gibidir:

- $I(32) - C(32, 5, 2) - C(32, 3, 1) - P(2) - FC(256) - FC(128) - FC(K)$
- $I(32) - C(64, 5, 2) - C(64, 3, 1) - P(2) - FC(512) - FC(256) - FC(K)$
- $I(32) - C(96, 5, 2) - C(96, 3, 1) - P(2) - FC(256) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(256) - FC(256) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(256) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(128) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(128) - FC(64) - FC(K)$
- $I(64) - C(64, 3, 1) - C(64, 3, 1) - P(2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - P(2) - FC(256) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(256) - FC(256) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(256) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(128) - FC(128) - FC(K)$
- $I(64) - C(64, 5, 2) - C(64, 3, 1) - P(2) - C(64, 3, 1) - C(64, 3, 1) - P(2) - FC(128) - FC(64) - FC(K)$

Buradaki K nesne kategori sayısını ifade etmektedir. $I(64)$, $64 \times 64 \times 64$ boyutlu hacimsel giriş verisini ifade etmektedir. Ayrıca bu modellerde $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ olmak üzere farklı parametreler ile $D(p)$ seyreltme işlemleri farklı katmanlarda denenmiştir. Görüleceği üzere kullanılan modele göre parametre sayısı oldukça daha fazla olan, daha karmaşık modeller de deneysel olarak araştırılmaktadır. Seçilen model, yukarıda anılan bileşenleri farklı parametre ve hiyerarşilerle bir araya getiren, yukarıda bazı örnekleri verilen birçok model arasından, deneysel olarak değerlendirilen alternatifler arasından seçilmiştir. Önerilen model, hem modeldeki parametre sayısı hem de deneylerde elde edilen doğruluk sonuçları dikkate alınarak seçilmektedir. Bu amaçla, bu model ile, [13]'te kullanılan modelin, giriş verisini daha iyi genellemesine yardımcı olması için kullanılan filtre sayısı iki katına çıkartılmaktadır. Ayrıca aşırıuyumlamayı azaltmak için yeni seyreltici katmanları ile tam-bağlantılı katmanları eklenmektedir. Önerilen mimarinin genel görünümü Şekil 4.2'deki gibidir.



Şekil 4.2. Önerilen yaklaşımın ağ mimarisi. Giriş katmanı, derinlik görüntülerinden elde edilen $32 \times 32 \times 32$ boyutlu hacimsel temsilleri kabul etmektedir. Evrişim katmanları, sırasıyla $5 \times 5 \times 5$ ve $3 \times 3 \times 3$ boyutlarındaki 64 adet filtreye sahiptir. Havuzlama katmanı, gelen hacimsel gösterimleri her uzamsal yönde 2 katsayı faktörü ile maksimum değerlerle düşürmektedir. Son tam bağlantılı katmanlar ise sırasıyla 128, 128 ve K (sınıf sayısı) birim içermektedir.

Kullanılan 3B ESA modelinin içerdiği bileşenlere ve bileşenlerin parametrelerine ilişkin detaylı bir görünüm, her düzeydeki hesaplanan parametre sayısı ile beraber Çizelge 4.1'de görülmektedir. Evrişim katmanlarının ilklendirilmeleri [30] yöntemi ile gerçekleştirilmektedir ve çıktı değerleri 0.01 parametresi ile *leaky* ReLU [70] aktivasyon fonksiyonundan geçirilmektedir. Havuzlama katmanında, maksimum havuzlama kullanılmaktadır. Tam-bağlantılı katmanlar, sıfır-ortalama Gaussian dağılımı ve $\sigma = 0.01$ parametresi ile ilklendirilmektedir. Son tam-bağlantılı katmanda, kategori sayısı ile eşdeğer K adet birim bulunmaktadır. Eğitim sırasında yitim değerini azaltmak için, stokastik gradyan düşüş (SGD) algoritması, $L2$ düzenleyicisi ve momentum optimizasyonu ile kullanılmaktadır. Düzenleme ve momentum parametreleri sırasıyla 0.001 ve 0.9'dur. Veriler yığınlara bölünerek ele alınırken, kullanılan yığın büyüklüğü 32'dir. Öğrenme

oranı (*learning rate*) 0.001 ile başlatılıp zaman içerisinde 10 kat azalacak bir politika ile yönetilmektedir. Ayrıca ESA eğitimlerinde yaygın bir alışkanlık olarak uygulanan verilerin çoğaltılması işlemi, örneklerin rastgele aynalanması ve kaydırılması ile elde edilen kopyaları ile sağlanmaktadır.

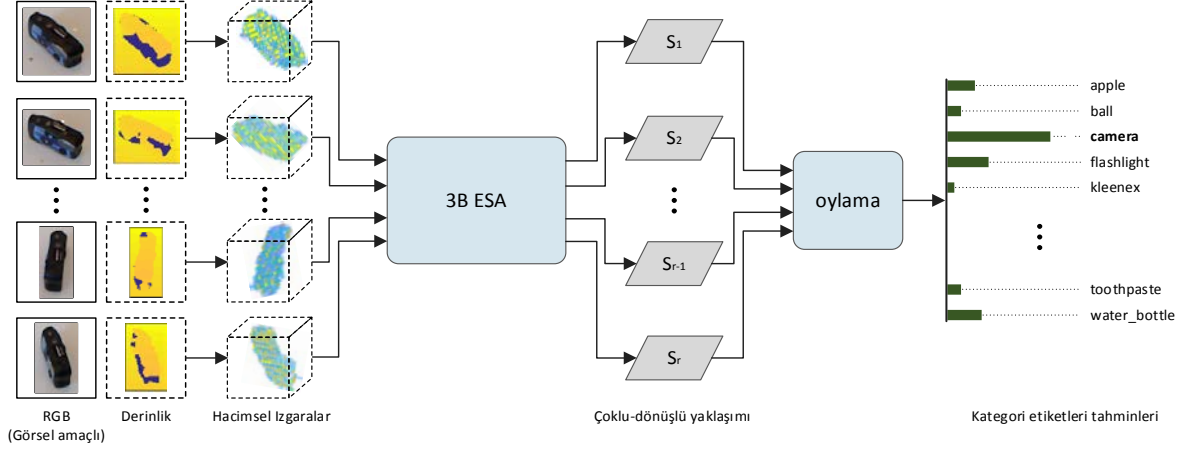
Çizelge 4.1. Kullanılan 3B ESA mimarisinin ayrıntıları.

Bileşen	Filtre Sayısı / Olasılık Faktörü	Kaydırma Adımı	Çıktı Boyutu	Parametre Sayısı
Giriş	–	–	$32 \times 32 \times 32$	–
Evrişim	$5 \times 5 \times 5$	2	$64 \times 14 \times 14 \times 14$	$64 \times 5 \times 5 \times 5 + 64$
Seyreltme	0.2	–	$64 \times 14 \times 14 \times 14$	–
Evrişim	$3 \times 3 \times 3$	1	$64 \times 12 \times 12 \times 12$	$64 \times (64 \times 3 \times 3 \times 3) + 64$
Havuzlama	$2 \times 2 \times 2$	2	$64 \times 6 \times 6 \times 6$	–
Seyreltme	0.3	–	$64 \times 6 \times 6 \times 6$	–
Tam-bağlantılı	–	–	128	$128 \times (64 \times 6 \times 6 \times 6) + 128$
Seyreltme	0.4	–	128	–
Tam-bağlantılı	–	–	128	$128 \times 128 + 128$
Seyreltme	0.5	–	128	–
Tam-bağlantılı	–	–	K	$K \times 128 + K$

4.3.3. Çoklu-Dönüştürme Yaklaşımı

Bu çalışmada, tek bir giriş görüntüsü ele alınmakla birlikte nesnelerin farklı açılardan çekilmiş görüntüleri, tam 3B modellerini sağlayamayan Kinect'in derinlik görüntülerini birleştirilerek, dönme değişmezliği ModelNet [26] kullanımına benzer bir şekilde elde edilmektedir. Bu amaçla, bir nesnenin birçok görünümünden elde edilen bilgileri derleyerek dönme değişmezliği sağlayan bir yaklaşım sunulmaktadır. Şekil 4.3, çoklu-dönüştürme yaklaşımının test aşamasını göstermektedir. Bu yaklaşım, dönme değişmezliğini sağlamak amacıyla eğitim sürecinde ağa eklenen dönme kopyaları ile ilişkilendirilebilir. Ayrıca, dönme görüntüleri boyunca filtrelerin yerel bağlantısını genişletmenin ve dönme görüntüleri boyunca ağırlıkların paylaşılmasının dolaylı bir yorumu olarak görülebilir. Çıkış katmanının aktivasyonları, test sürecinde dönme görüntüleri üzerinde toplanıp

birleştirilmektedir. Ağ, nesnelerin çok sayıda dönme kopyalarını alır ve nesne sınıfının tanımlanması için skor değerlerinde son bir oylama yaklaşımı gerçekleştirilir. Algoritma 4.1, test zamanında çoklu-dönüştü yaklaşımını kullanarak bir nesne kategorisinin tanınmasını, özetle açıklamaktadır.



Şekil 4.3. Önerilen çoklu-dönüştü nesne tanıma yöntemi. 3B ESA, nesnelerin çoklu dönüşlerini giriş olarak kabul edip, olası nesne kategorileri için S_r skor değerlerini üretir. Daha sonra nesne kategorisi, nihai bir oylama işlemi ile atanmaktadır. Sol sütundaki RGB görüntüleri görsel amaçlıdır. Önerilen yöntem, sadece derinlik görüntülerini kullanır.

Algoritma 4.1. Çoklu-dönüştü derinlik verilerini kullanarak nesne kategorilerini tanıma.

```

giriş :  $I$  dönme görüntüleri,  $C$  kategoriler
çıkış : kategori etiketi tahmini

for her bir  $I_r \in I$  do
     $S_r \leftarrow 3B - ESA(I_r)$  { $I_r$  için skor değerlerini al}
end
for her bir kategori  $c \in C$  do
     $F_c \leftarrow 0$  {kategori  $c$ 'nin nihai skorunu tutar}
    for her bir dönme  $I_r \in I$  do
         $F_c \leftarrow F_c + S_c(r)$ 
    end
end

return  $F$  üzerinden maksimum skorlu kategori  $c$ 

```

4.4. Deneysel Değerlendirmeler

Önerilen yaklaşım iki standart veri kümeleri olan Washington RGB-D [67] ve 2D3D Nesne [80] veri kümelerinde değerlendirilmiştir. Bu bölümde, ilk önce veri kümeleri deney düzenekleriyle açıklandıktan sonra deneysel sonuçlar analizlerle değerlendirilmektedir. Daha sonra, önerilen yaklaşımların performansı, hem Washington RGB-D hem de 2D3D Nesne veri kümeleri üzerinde derinlik bilgisi kullanan önceki yöntemlerle karşılaştırılmaktadır. Diğer yöntemlerin sonuçları orijinal yayınlardan alınmıştır. Son olarak, bu tez metni kapsamında, hacimsel giriş görüntülerini dönme matrisleri ile çoğaltan yaklaşımdan, derinlik bilgisinin yanı sıra renk bilgisinin de hesaba katıldığı yaklaşımlardan ve bunların deneysel sonuçlarına ilişkin değerlendirmelerden bahsedilecektir.

4.4.1. Veri Kümeleri ve Kurulumları

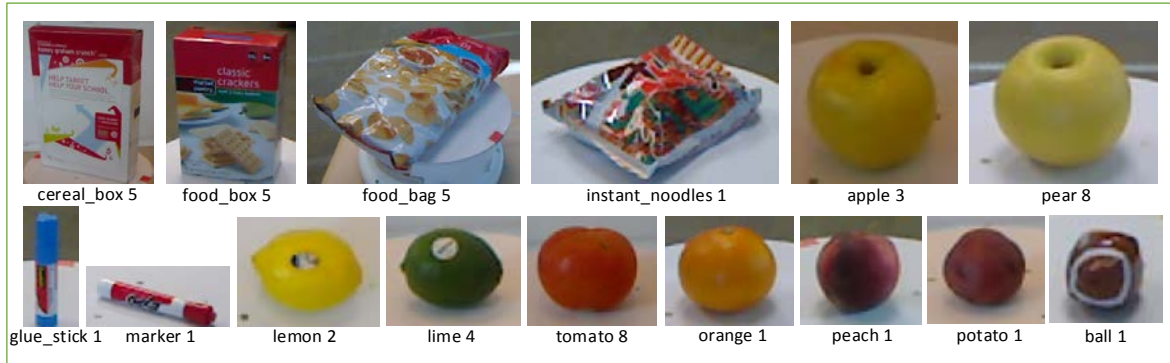
4.4.1.1. Washington RGB-D Veri Kümesi

Washington RGB Nesne veri kümesi, 300 alt kategori örnekleminde (*instance*) toplam 207,662 görüntüye sahip 51 nesne kategorisini içermektedir. Her alt kategori örneklemini, nesnelerin, üç yükseklik açısından (30° , 45° , 60°) kaydedilmiş üç farklı video dizilerinin görüntülerini içermektedir. Bu veri kümesinin RGB-D nesne tanımada yaygın olarak kullanılmasının nedenleri, hem tanımlanmış en kapsamlı ve eski veri kümesi olmasından hem de bu veri kümesinin zorlu bir veri kümesi olmasından dolayıdır. Giriş bölümünde anlatılan nesne tanıma problemini zorlaştıran unsurlar bu veri kümesinde görülmektedir: (i) Veri kümesi, zengin bir sınıf-içi çeşitliliğe sahiptir. Şekil 4.4'te “top” (*ball*) ve “kahve kupası” (*coffee mug*) nesne kategorilerinin örneklemlerinden, bu çeşitliliği gösteren görüntü örnekleri görülmektedir. (ii) Veri kümesi, sınıflar-arası benzerliği yüksek, farklı nesne kategorilerinde benzer örneklemler içermektedir. Şekil 4.5, sınıflar-arası benzerliği yüksek örnekleri göstermektedir. Bu veri kümesinde, şekilsel olarak birbirlerine benzeyen birçok nesne kategorisi mevcuttur (örn. domates (*tomato*), patates (*potato*), top (*ball*), portakal (*orange*), şeftali (*peach*), elma (*apple*) kategorileri gibi). Bu durum kategorilerin kolayca karıştırılmasına yol açabilir. Öte yandan bu çalışmada sadece derinlik bilgisinin kullanılması göz önüne alındığında, renk bilgisinden yoksun sadece geometrik yapılarına göre bu kategorileri ayırt etmek daha da zorlaşmaktadır. (iii) Bakış açısı ve ölçek çeşitliliği ile nesnelerin sadece bir kısmının olduğu görüntü örnekleri, özellikle derinlik bilgisinin zayıf kaldığı örneklerin çokluğu, görüntü gürültüleri ve bozulmalar, bu veri kümesinde nesne kategorilerini tanımayı zorlaştıran diğer etmenlerdir. Örneğin cam gibi parlak yüzeylerden yansıyan ışıklar, “hesap makinesi” ya da “kamera” kategorileri için ve “dış

fırçası” kategorisi gibi örnekler için de ince boyutlarından dolayı, eksik derinlik bilgilerinden dolayı tanıma zorlaşmaktadır.



Şekil 4.4. Washington RGB-D Nesne veri kümesinin sınıf-içi çeşitliliğini gösteren alt kategori örneklerinden görüntü örnekleri. “Top” ve “Kahve Kupası” sınıfları için farklı örneklerden görüntüler.



Şekil 4.5. Washington RGB-D Nesne veri kümesinde sınıflararası benzerliği gösteren örnekler. Her görüntü farklı bir nesne kategorisine aittir.

Hacimsel temsiller, farklı senaryolarda değerlendirilirken (Bölüm 4.4.2.1), tüm veri kümesi rastgele üç kısma ayrılarak ele alınmaktadır: %60 eğitim, %20 doğrulama ve %20 test bölmeleri. Bu deneyler için model, 120 epoch’a kadar eğitilmektedir.

Bu veri kümesindeki diğer çalışmalarla kıyaslama yapılırken, [67] çalışmasındaki deney düzeneğini izlenerek, veri kümesindeki derinlik görüntüleri 1/5 oranında alınarak, yaklaşık 41,500 görüntüye sahip olacak şekilde veri kümesi örneklenmektedir. Bu kesimde, Bölüm 4.4.2.1’de bulunan en iyi gösterim senaryosu ve nesnelerin yalnızca tek bir dönme görüntüsü ile sonuçlar alınmaktadır. Doğruluk performansı hesaplanırken, Lai ve diğerlerinin [67] kullandıkları 10 eğitim/test bölmeleri kullanılmaktadır ve bu 10 bölmenin ortalama sonucu raporlanmaktadır. Her bölmede, bir kategori için, bir alt-

kategori örnekleme (*instance*) test için seçilirken, geri kalan örneklemler eğitim için kullanılmaktadır. Bu deneyler için model, 820 epoch'a kadar eğitilmektedir.

4.4.1.2. 2D3D Nesne Veri Kümesi

2D3D Nesne veri kümesi, 14 nesne kategorisini ve bunların altında da toplamda 155 alt kategori örnekleme içermektedir. Her örnekleme, aynı nesnenin 10°'lik artan açılarla kaydedilmiş toplam 36 örnek görüntüsü vardır. Bu veri kümesini kullanırken, veri kümesine uygun olarak [80] çalışmasındaki deney ayarları takip edilmektedir. Buna göre, veri kümesindeki görüntüler, birer atlanarak 1/2 oranında örneklenmektedir. Böylece toplamda 2790 görüntü örneği kullanılmaktadır. Daha sonra, eğitim için kategori başına 6 örnekleme rastgele seçilir ve geri kalan örneklemler test için kullanılır. "Makas" (*Scissor*) kategorisi 6'dan az örnekleme içeren tek kategoridir. Bu kategori için 4 örnekleme eğitimde kullanılmakta ve test için, geri kalan 1 örnekleme kullanılmaktadır. Böylece, eğitim kümesi toplamda 1476 görüntüyü içeren 82 örneklemden oluşurken, test kümesi 1314 görüntüyü içeren 73 örneklemden oluşmaktadır.¹ 2D3D Nesne veri kümesi, Washington RGB-D Nesne veri kümesi kadar büyük değildir. Derin öğrenme yöntemleri, eğitimde fazla veri gerektiren genel olarak veriye aç modellerdir. Bu nedenle, bu veri kümesinde derin öğrenme ile nesne tanımanın temel zorluğu, veri kümesinin küçük boyutudur. Bu veri kümesi ile yapılan deneylerde model 5000 epoch'a kadar eğitilmektedir.

4.4.1.3. Çoklu-Dönümlü Kurulumu

Yukarıda belirtilen kurulumlar, nesne kategorilerini tanımak için tek bir derinlik görüntüsünü (tek-dönümlü) kullanırlar. Bu çalışmada ayrıca çoklu-dönümlü yaklaşımına uygun olarak, veri kümeleri ayarlanmaktadır. Washington RGB-D Nesne veri kümesi, farklı yükseklik açılarından çekilmiş görüntüleri karışık bir düzende içermektedir. Her bir örnekleme için 30°, 45° ve 60° açılardan çekilmiş görüntü dizileri, görüntülerin dönme açıları bakımından birbirlerini takip edecek şekilde düzene getirilerek ve her örnek için toplamda bir nesnenin tam dönüş görüntülerini içerecek şekilde düzene sokulmaktadır. Oluşturulan bu yeni düzendeki veri kümesi Washington RGB-D veri kümesinin bir alt kümesini meydana getirmektedir. 30°, 45° ve 60° dizilerinin her biri, tek bir nesnenin farklı dönme açıları ile çekilmiş görüntülerini içermektedir. Bu yüzden, her bir dizi için sezgisel olarak 20°'lik açı farkıyla olacak şekilde toplamda 18 dönme görüntüsü

¹ Browatzki ve diğerleri [80], eğitim ve test için bölmeleri sırasıyla 82 ve 74 olarak ifade etmektedirler. Ancak, <http://www.kyb.mpg.de/~browatbn> bağlantısında toplamda 155 örnekleme mevcuttur. İlgili yazar, *Cup* kategorisinde 14 alt kategori örnekleme yerine 13 örnekleme ve toplamda 155 kategori örnekleme olduğunu teyit etmiştir.

seçilmektedir. Böylece bu yeni veri kümesi toplamda 16,200 görüntü içermektedir. Daha sonra, [67]'deki deney düzeneğine uygun olarak test gerçekleştirmek için, her bir kategoriden bir alt kategori örnekleme test için ayrılırken, geri kalan $300 - 51 = 249$ örnekleme eğitim için kullanılmaktadır. Deneyler, yine Lai ve diğerlerinin [67] sağladıkları 10 eğitim/test bölmelerinde yapılarak, bunların ortalama doğruluk değerleri raporlanmaktadır. Bu deneylerde de, Washington RGB-D Nesne veri kümesinin diğer deneylerinde olduğu gibi model 820 epoch'a kadar eğitilmektedir.

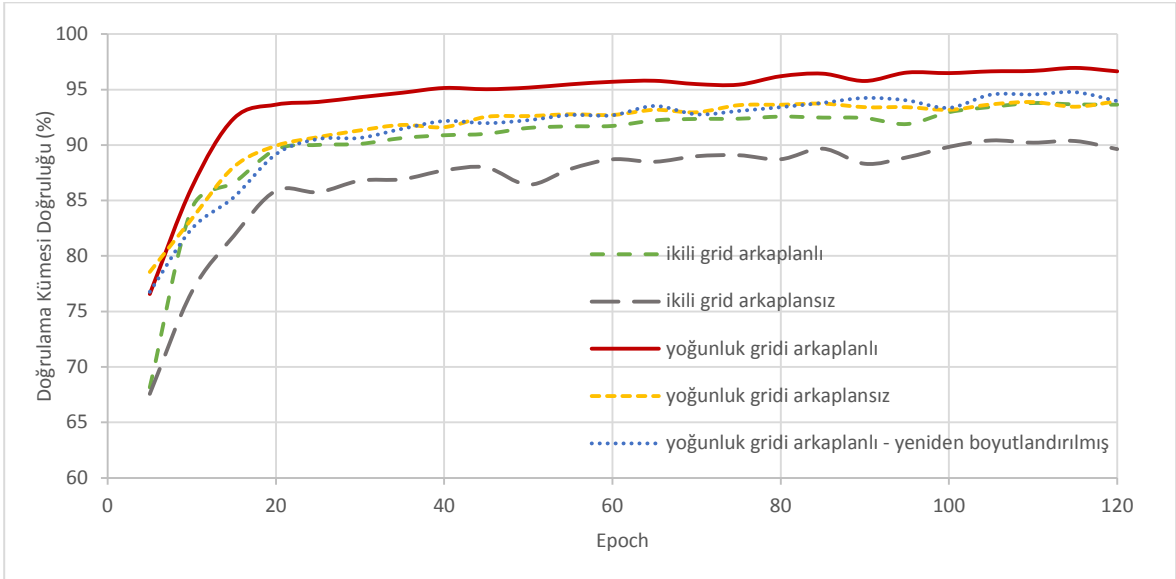
2D3D Nesne veri kümesinde, her bir alt kategori örnekleminde nesne görüntüleri 10° lik açılarla kaydırılarak toplamda 36 görüntü şeklinde düzenli bir yapıda verilmektedir. Bu yüzden bu veri kümesinde, çoklu-dönüştürme yaklaşımı uygulanırken, veri kümesi olduğu gibi kullanılmaktadır. Bölüm 4.4.1.2'de anlatıldığı gibi her örnekleme birer görüntü atlanarak 20° lik açı farkı ile toplamda 18 görüntü örneği alınmaktadır. Yine tek-dönüştürme yaklaşımda olduğu gibi, [80] çalışmasındaki deney ayarları takip edilerek, model 5000 epoch'a kadar eğitilmektedir.

4.4.2. Sonuçlar

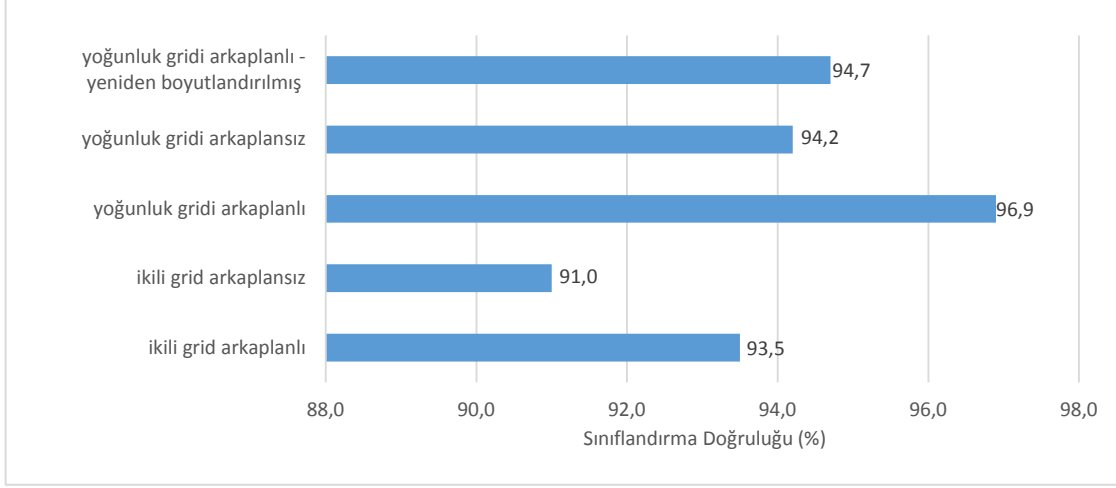
4.4.2.1. Hacimsel Grid Farklılıkları

İlk olarak, Washington RGB-D Nesne veri kümesini kullanarak nesne maskelerinin etkilerini görmek için, maskelerin kullanıldığı (arkaplanlı) ve kullanılmadığı (arkaplanlı) ikili ve yoğunluk hacimsel gridleri değerlendirilmektedir. Ayrıca, görüntülerin yeniden boyutlandırılmalarından dolayı meydana gelebilecek potansiyel kırpma (*cropping*) ve eğrilme (*warping*) durumlarının etkilerini incelemek için, giriş görüntülerinin yeniden boyutlandırılarak ele alındığı deneyler yapılmaktadır. Şekil 4.6 ve Şekil 4.7, bu kapsamda yapılan deneylerin sırasıyla doğrulama kümesindeki doğruluğunu ve test performansını içeren sonuçlarını göstermektedir. Hacimsel yoğunluk gridinin ikili gridden daha iyi performans gösterdiği görülmektedir. Bu durum, bir vokseldeki nokta yoğunluğunun dikkate alındığı temsiline, bir voksele yansıtılan bir noktanın varlığını/yokluğunu dikkate alan ikili grid yapısına göre, daha iyi bir sınıflandırma performansı için daha fazla bilgi verdiğini doğrulamaktadır. Ayrıca, hiçbir nesne maskesinin kullanılmadığı deneyleri yaparak, önerilen yaklaşımın arkaplan karmaşasını ele alabildiği de gösterilmektedir. Beklentilerin aksine, nesne maskelerinin kullanılmasının sınıflandırma performansını olumsuz yönde etkilediği gözlemlenmektedir. Bu durumun bir sebebi, veri kümesi ile sağlanan nesne maskelerinin yeterince iyi olmamaları olabilir. Son olarak, ESA çalışmalarında eşit boylu giriş görüntülerinin gereksiniminden ötürü, yapılan yeniden

boyutlandırma (*resize*) işleminin etkisi incelenmektedir. ESA'nın sabit ve eşit boyutlu giriş görüntülere olan gereksinimleri, [64] çalışmasında yazarların değindikleri kırpmalardan ve eğrilmelerden kaynaklı performans kayıplarına neden olabilir. Bu amaçla hacimsel grid temsilleri, yeniden boyutlandırılarak elde edilen 148×148 boyutlu derinlik görüntülerinden oluşturulmaktadır. Şekillerde de görüldüğü gibi, sınıflandırma performansı beklendiği gibi düşmektedir. Giriş görüntülerinin orijinal boyutlarından elde edilen hacimsel temsillerin, yeniden boyutlandırılarak elde edilen temsillere göre daha iyi performans gösterdiği görülmektedir. Bu durum, hacimsel temsilleri elde ederken nokta bulutunun yansıtılması yaklaşımının, sabit boylu giriş görüntüleri gereksinimi sınırını aşarak, olası performans kayıplarını önlediğini göstermektedir. Dolayısıyla, sonuç olarak bu bölümde yapılan deneylerde, yoğunluk gridinin görüntülerin arkaplanlarını kaldırmaksızın kullanıldığı senaryosunda en iyi performansın elde edildiği gözlenmektedir. Deneylerin geri kalan kesimlerinde, en iyi performans elde edilen hacimsel yoğunluk grid temsilleri arkaplan kombinasyonu ile kullanılmaktadır.



Şekil 4.6. Hacimsel gridlerin çeşitli parametrelerle Washington RGB-D Nesne veri kümesinde doğrulama kümesi üzerinde etkileri.



Şekil 4.7. Hacimsel gridlerin çeşitli parametrelerle Washington RGB-D Nesne veri kümesi test kümesindeki doğruluk performansları.

4.4.2.2. Washington RGB-D Nesne Veri Kümesinde Karşılaştırma

Washington RGB-D Nesne veri kümesindeki sonuçların karşılaştırılması Çizelge 4.2’de raporlanmaktadır. Hem tek-dönüslü hem de çoklu-dönüslü yaklaşımlarda aynı deney kurulumları ile 10 eğitim/test bölmeleri kullanılmasına rağmen, çoklu-dönüslü yaklaşım, bu yaklaşıma uygun hale getirmek için Washington RGB-D Nesne veri kümesinden oluşturulan ve bu veri kümesinin bir alt kümesi olan verilere uygulanmaktadır. Bundan dolayı çoklu-dönüslü yaklaşım sonucu, çizelgede bu durumu belirtmek amacıyla * belirteci ile verilmektedir. Ayrıca, önerilen bu yaklaşım, veri kümesinde çoklu giriş görüntülerini ele alan tek yöntemdir. Bu nedenle sonuç, çizelgede ayrı bir satırda verilmektedir. Çoklu-dönüslü yaklaşım, veri sayısı küçülmesine rağmen dönme değişmezliği sağlayarak tanıma performansını önemli ölçüde artırmaktadır. Tek-dönüslü yaklaşımda ise, görüntülerin elde edildiği bakış açılarına bağlı eksik hacimsel temsillere rağmen, ilgili diğer çalışmalarla rekabetçi sonuçlar verdiği görülmektedir ve CNN-SPM-RNN [64] ile Hypercube [83] çalışmaları dışındaki diğer tüm yöntemleri geride bırakan bir performans sergilemektedir. Zaki ve diğerlerinin yöntemi [83], farklı veri türlerini beraber kullanarak en iyi sonucu elde etmektedir. Ancak, bu yöntemde önerilen saklı (*embedded*) nokta bulutu temsilleri, renk bilgilerinden yararlanmaktadır. Bu nedenle bu yöntem, diğer yöntemlerin aksine sadece derinlik bilgisini kullanmamaktadır. Çizelgede bu durumu belirtmek amacıyla ilgili sonuç satırında, † belirteci kullanılmıştır. Sadece derinlik görüntülerinin kullanıldığı izole deneyleri, 79.4% tanıma doğruluğuna erişmektedir. Öte yandan, Cheng ve diğerlerinin önerdikleri CNN-SPM-RNN [64] yöntemi sadece derinlik görüntülerine değil, aynı zamanda yüzey normallerine de ayrı olarak uygulandıktan sonra elde edilen nitelikler,

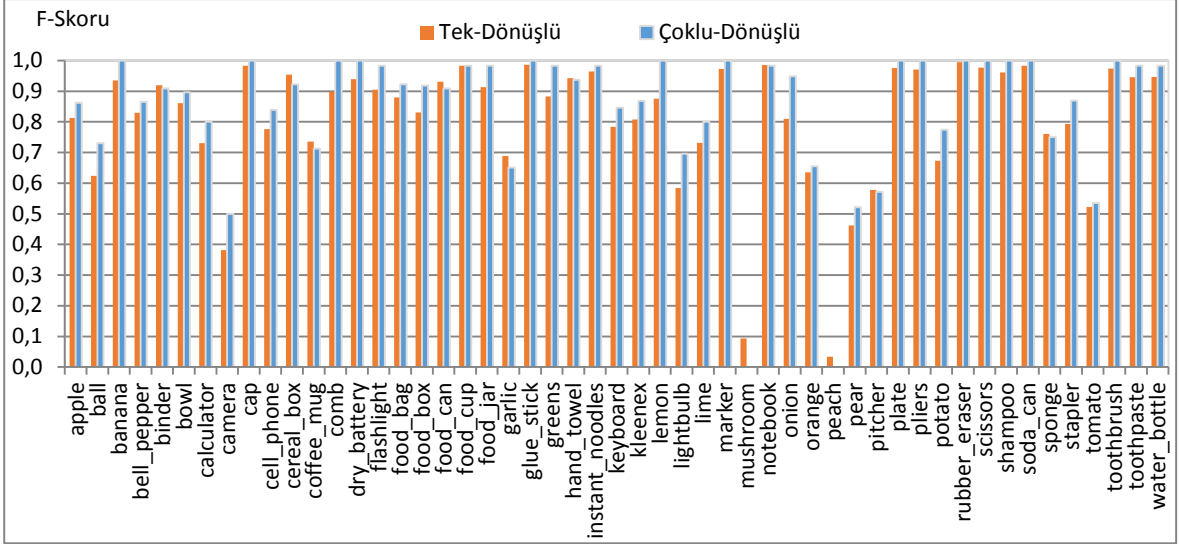
nihai derinlik başarısını hesaplamak üzere birleştirilerek değerlendirilmektedir. Bu duruma karşın, bu çalışmada önerilen yöntemde sadece ham derinlik görüntüleri kullanılmaktadır.

Çizelge 4.2. Washington RGB-D Nesne veri kümesinde derinlik verilerini kullanan ilgili çalışmaların kategori tanıma doğruluğu karşılaştırması.

Yöntem	Doğruluk (%)
Kernel SVM [67]	64.7 ± 2.2
HKDES [81]	75.7 ± 2.6
SSL [63]	77.7 ± 1.4
KDES [72]	78.8 ± 2.7
CNN-RNN [56]	78.9 ± 3.8
HMP [62]	81.2 ± 2.3
Subset-RNN [61]	81.8 ± 2.6
Volumetric [13]	82.0 ± 2.3
CNN-SPM-RNN [64]	83.6 ± 2.3
Hypercube [83] [†]	85.0 ± 2.1
Bu çalışma (tek-dönüslü)	82.4 ± 2.2
Bu çalışma (çoklu-dönüslü) *	85.9 ± 2.9

Şekil 4.8’de, tek-dönüslü ve çoklu-dönüslü yaklaşımlarının Washington RGB-D Nesne veri kümesindeki her bir nesne kategorisi için ayrı ayrı elde ettikleri f-skorumları gösterilmektedir. Buna göre, birden çok dönüş görüntülerinin kullanılması, çoğunlukla sonucu iyileştirmektedir. Genel olarak, hem tek-dönüslü hem de çoklu-dönüslü yaklaşımlarda düşük sonuçlara sahip birkaç nesne kategorisi vardır. Bu kategoriler, *camera*, *muhroom* ve *peach* nesne kategorileridir. Bu nesne kategorilerinin ortak sorunu, veri kümesindeki en az alt kategori örneklem sayısı olan sadece üçer tane örneklem içermeleridir. Bu sayı, 14 örnekleme kadar alt kategorileri içeren nesne sınıfları göz önüne alındığında oldukça küçük kalmaktadır. Kategori örneklemelerindeki bu dengesiz dağılım, az sayıdaki örnekleme sahip nesne sınıflarının tekil başarılarını düşürdüğü gibi sistemin genel başarısını da düşürmektedir. Çünkü veri kümesindeki bu dengesizlik, daha fazla sayıda örnekleme sahip nesne kategorilerinin lehine öğrenmeyi saptırır. Hatta *muhroom* ve *peach* kategorileri için veri miktarının iyice azaldığı çoklu-dönüslü yaklaşımda sonuç daha da kötü olmaktadır. Bunların yanı sıra, kötü performansın diğer nedenleri veri kümesindeki

sınıf-içi çeşitlilik ve sınıflar-arası benzerliktir. Öte yandan, kamera gibi parlak yüzeylere sahip örnekler derinlik bilgilerini bozabilir, çünkü derinlik algılayıcıları bu yüzeylerden yansımaları doğru şekilde ele alamaz.



Şekil 4.8. Washington RGB-D Nesne veri kümesindeki her bir kategori için f-skorları.

Washington RGB-D Nesne veri kümesindeki yanlış sınıflandırılmış nesne kategorileri örnekleri Şekil 4.9’da sunulmuştur. Şekildeki ilk sütun (a), örneklerin hacimsel temsillerini göstermektedir. Burada hem nesnelere hem de arkaplanın beraber görülmektedir. İkinci sütun (b), bu örneklerin RGB görüntülerini göstermektedir. Son sütunda (c) ise, sistemin yanlış öngördüğü nesne kategorilerinden birer örnek gösterilmiştir. Şekildeki RGB görüntüleri (b ve c sütunları), görsel amaçlı kullanılmıştır. İlk sütunda sergilendiği gibi nesne temsillerinde, sadece derinlik görüntülerinden elde edilen hacimsel gridler kullanılmaktadır. Şekilde görülebileceği gibi, yanlış sınıflandırılmış nesnelerin geometrik yapıları aslında birbirleriyle oldukça benzerdir. Bu durum yanlış sınıflandırılmalarının temel sebebidir. Yukarıdan aşağıya doğru, “mantar” (*mushroom*) “top” (*ball*) ile, “şeftali” (*peach*) “portakal” (*orange*) ile, “şeftali” (*peach*) “elma” (*apple*) ile, “armut” (*pear*) “elma” (*apple*) ile, “domates” (*tomato*) “patates” (*potato*) ile ve “kamera” (*camera*) “sünger” (*sponge*) ile karıştırılmıştır.



Şekil 4.9. Washington RGB-D Nesne veri kümesindeki yanlış sınıflandırılmış kategori örnekleri. RGB görüntüler görsel amaçlı kullanılmıştır. Önerilen çalışmada sadece derinlik görüntüleri kullanılmıştır. (a) Karıştırılan nesnenin hacimsel temsili. (b) Gerçek-referans örneğinin RGB görüntüsü. (c) Yöntemin yanlış öngördüğü nesne kategorisine ait örnek bir RGB görüntü.

4.4.2.3. 2D3D Nesne Veri Kümesinde Karşılaştırma

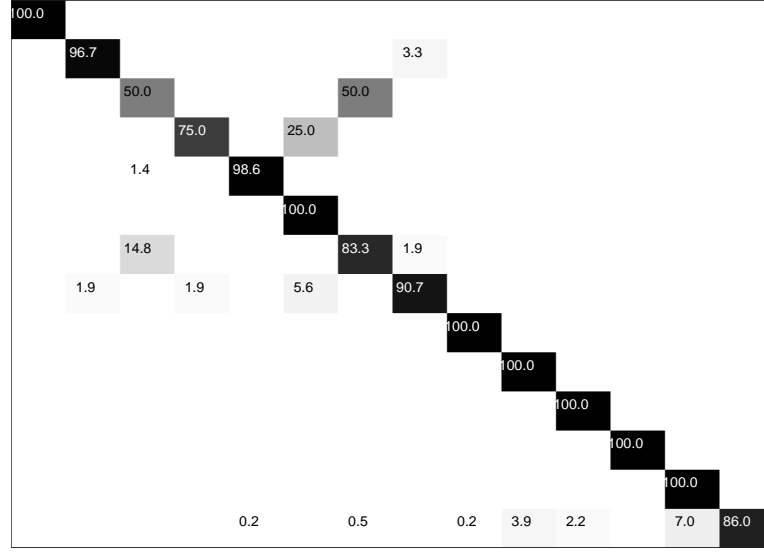
Çizelge 4.3'te önerilen yöntemin, 2D3D Nesne veri kümesini kullanan önceki çalışmalarla olan sınıflandırma doğruluk karşılaştırması gösterilmektedir. Bölüm 4.4.1'de anlatıldığı gibi, hem tek-dönüslü hem de çoklu-dönüslü yaklaşımları aynı kurulumlarla bu veri kümesinde uygulanmaktadır. 2D3D Nesne veri kümesinin küçük boyutuna rağmen, önerilen yöntem her iki yaklaşım için de önemli bir performans sergilemektedir. Çoklu-dönüslü yaklaşım, bir nesnenin birden çok dönme görüntüsünden elde edilen bilgileri birleştirerek en iyi sonuca ulaşmaktadır. Ancak, yukarıda bölüm 4.4.2.2'de anlatıldığı gibi, önerilen çoklu-dönüslü yaklaşım, birden fazla giriş görüntülerinden yararlanan tek yöntemdir. Bu nedenle, bu yaklaşımın sonucu, çizelgede ayrı bir satırda verilmektedir. Öte yandan, önerilen tek-dönüslü yaklaşım, Subset-RNN [61] ile beraber Hypercube [83] yönteminden sonra en yüksek tanıma doğruluğunu vermektedir. Hypercube [83] yönteminde, kullanılan saklı nokta bulutu temsillerinde renk bilgisinden yararlanıldığından dolayı, bu yöntemin sonucu, bu durumu bildirmek üzere diğerlerinden ayrı olarak çizelgede, [†] belirteci ile verilmektedir. Dolayısıyla, bu yöntemin fazladan 1.5% başarı oranı anlaşılabilir düzeydedir.

Çizelge 4.3. Derinlik verilerini kullanan ilgili çalışmaların 2D3D Nesne veri kümesinde kategori tanıma doğruluğu karşılaştırması.

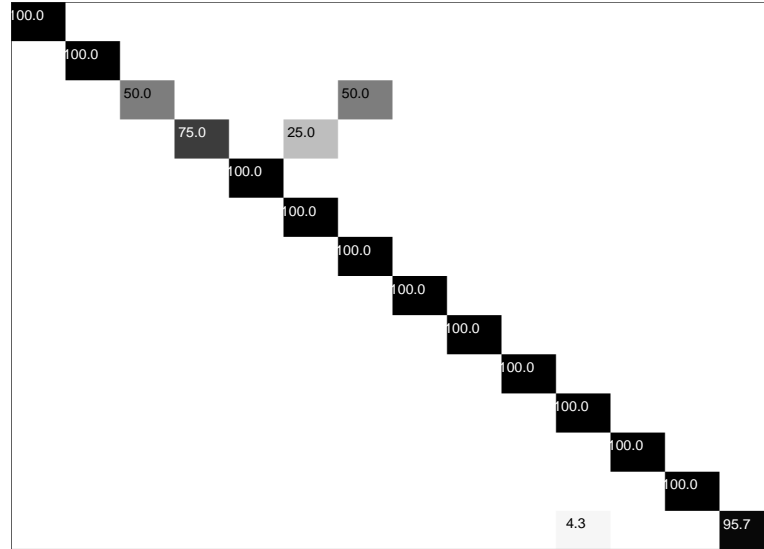
Yöntem	Doğruluk (%)
Browatzki vd. [80]	74.6
HMP [62]	87.6
CNN-SPM-RNN [64]	89.4
Subset-RNN [61]	90.2
Hypercube [83] [†]	91.6
Bu çalışma (tek-dönüslü)	90.1
Bu çalışma (çoklu-dönüslü)	94.5

2D3D Nesne veri kümesinin 14 kategorisi üzerindeki, tek-dönüslü ve çoklu-dönüslü yaklaşımlarının hata matrisleri sırasıyla Şekil 4.10 ve Şekil 4.11'de gösterilmektedir. Şekillerden görülebileceği üzere, çoğu kategori doğru şekilde sınıflandırılmaktadır. Hata matrislerindeki sonuçlar incelendiğinde, “Şişe” (*Bottle*) ve “Konserve Kutuları” (*Cans*) kategorileri en yüksek hata oranına sahiptir. “Şişe” (*Bottle*) örnekleri “Bulaşık Sıvısı” (*DishLiquid*) ile karıştırılırken, “Konserve Kutuları” (*Cans*) ise “Kupa” (*Cup*) örnekleri ile

kariřtirilmektedir. Washington RGB-D Nesne veri kümesinde olduđu gibi, bu yanlış sınıflandırmalar da temel olarak nesnelerin Őekil benzerliklerinden kaynaklanmaktadır.



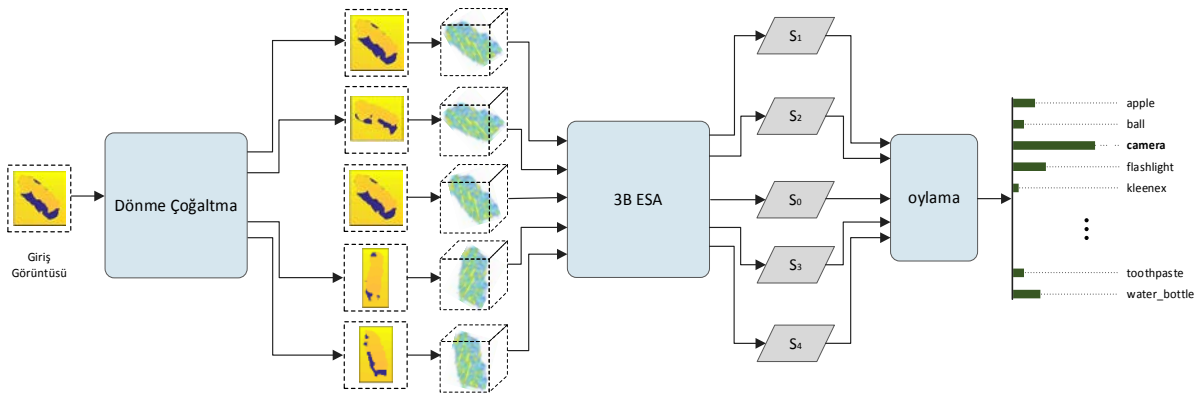
Őekil 4.10. Tek-dönüřlü yaklaşımın, 2D3D Nesne veri kümesindeki hata matrisi.



Őekil 4.11. 2D3D Nesne veri kümesinde, çoklu-dönüřlü yaklaşımının hata matrisi.

4.4.3. Çoklayan-Dönüşlü Yaklaşımı ve Deneysel Sonuçlar

Bölüm 4.3.3'te anlatılan çoklu-dönüşlü yaklaşımda, birden fazla giriş görüntüsü beraber değerlendirilmektedir. Bu amaçla, kullanılan veri kümelerinde aynı nesnenin farklı açılardan çekilmiş görüntüleri kullanıldığında, nesnenin birçok yönünden elde edilen dönme görüntülerinin ipuçları, nesnenin bütününe daha iyi ifade etmektedir. Böylece Kinect algılayıcısı ile sadece bir açıdan çekilmiş ve nesneye ilişkin tam bir model vermeyen yapının dezavantajı aşılmış olmaktadır. Öte yandan, anlatılan bu yaklaşımın yanı sıra giriş görüntülerini dönme matrisleri ile çoklayarak benzer mantıkla kullanmak da mümkündür (Şekil 4.12). Elimizde nesnenin tüm yönlerinden çekilmiş bütün bir model olmuş olsaydı, bu yaklaşımla başarı derecesini önemli derecede artırmak mümkün olabilirdi. Kinect görüntüleri, nesnelerin kameraya dönük yönlerinden bilgiler sağladıklarından ötürü, çoklayan-dönüşlü yaklaşımı, bu görüntülerde kullanılırken dönme açılarının iyi seçilmesi gerekir.



Şekil 4.12. Çoklayan-dönüşlü yaklaşımla nesne tanıma. Giriş görüntüsü saat yönü ve saat yönünün tersinde belli açılarla z-ekseni boyunca döndürülerek çoğaltılmaktadır. Elde edilen görüntüler, nesnenin farklı açıdan görünümünü sağlamaktadır.

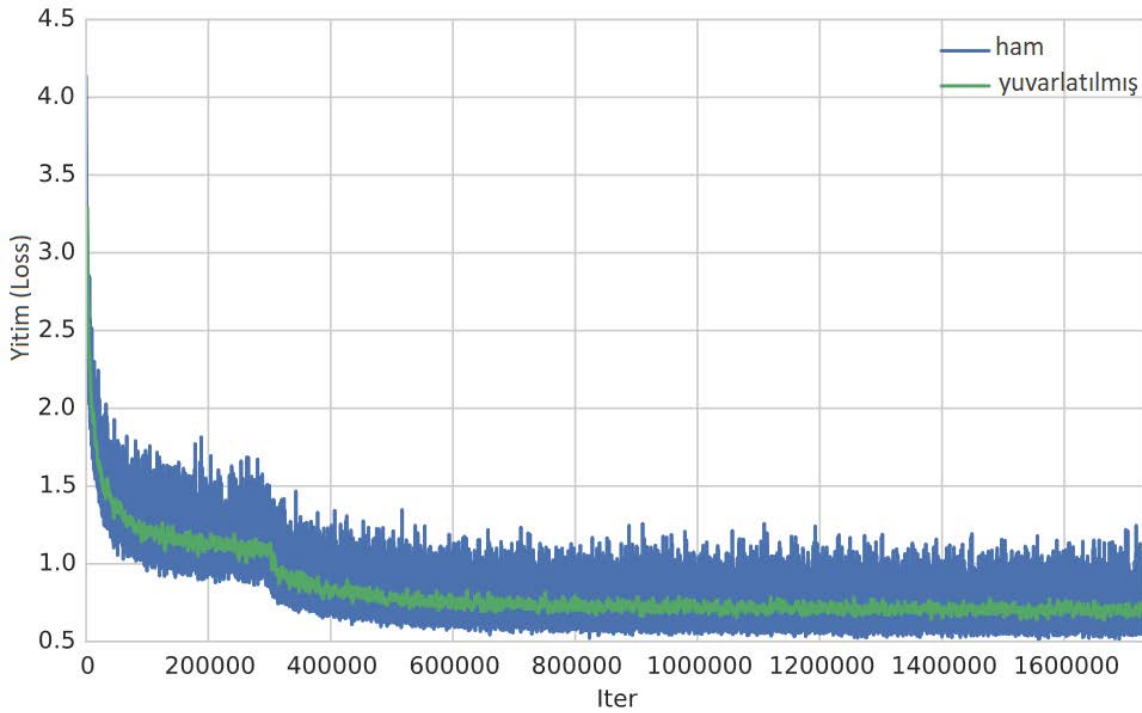
Verilen bir $P \in \mathbb{R}^{m \times 3}$ nokta bulutu, $R \in \mathbb{R}^{3 \times 3}$ dönme matrisi ile çoğaltılarak elde edilen nokta bulutlarından hacimsel gridler elde edilmektedir. Daha sonra, oluşturulan temsiller çoklu-dönüşüm yaklaşımı gibi (Bkz. Bölüm 4.3.3) ele alınmaktadır. Dönme matrisi R , z-ekseni etrafında saat yönü ve saat yönünün tersinde olmak üzere aşağıdaki gibi kullanılmaktadır.

$$R = \left\{ \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}$$

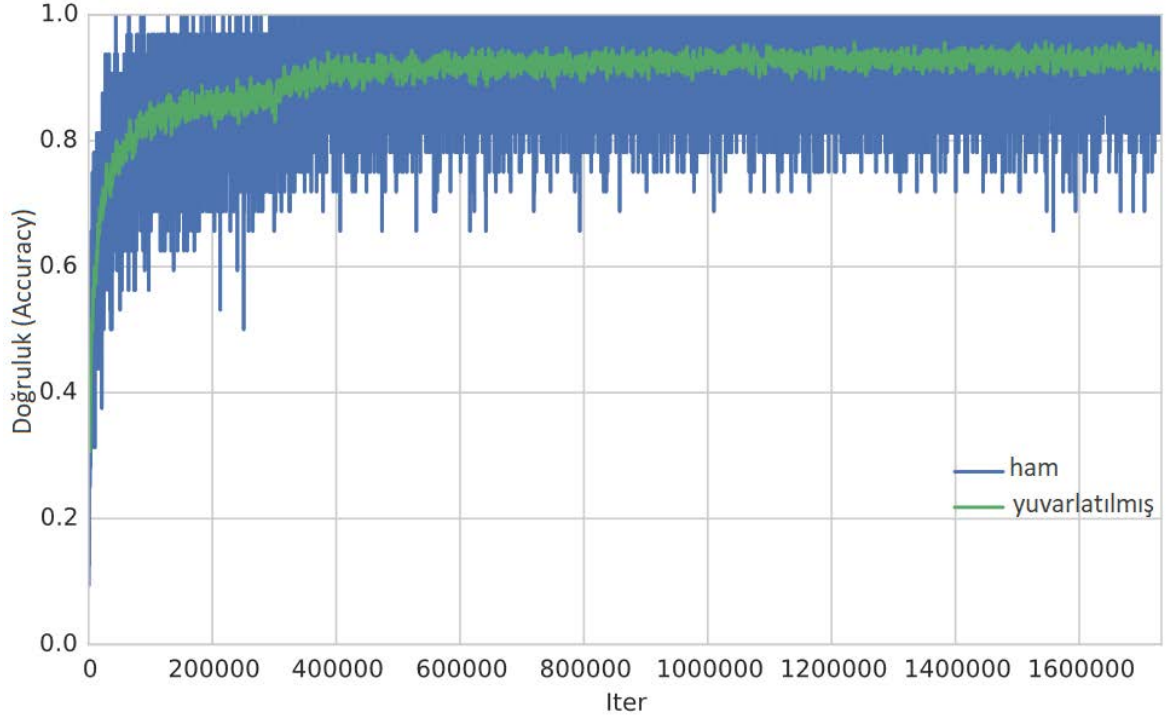
Deneysel olarak, z-ekseninin her iki yönünde 15° açılarla kayırarak orijinal görüntü ile beraber 5 tane dönme görüntüsünün, en etkili sonuç verdiği tespit edilmektedir.

Dolayısıyla $\theta = \{-30^\circ, -15^\circ, 15^\circ, 30^\circ\}$ kullanılmaktadır. Bu açılar mantıklıdır, çünkü daha büyük açılar nesnelerin boş olan arka taraflarını gösterirken, daha küçük açılar ise orijinal görüntünün pek değişmemesi demektir.

Çoklayan-dönüştürme yaklaşımı için deneyler, yine Washington RGB-D [67] ve 2D3D Nesne [80] veri kümelerinde yapılmıştır. Her iki veri kümesi için de deneyler, orijinal makalelerde önerildiği ayarlamalar ile yapılmıştır. Washington RGB-D Nesne veri kümesi için sağlanan 10 eğitim/test bölmesi kullanılmıştır ve sonuç bu 10 bölmenin ortalaması alınarak standart sapması ile beraber raporlanmıştır. Washington RGB-D Nesne veri kümesinde, 3B ESA modeli 320 epoch'a kadar eğitilmektedir. Veri boyutunun dönme görüntüleri ile beraber $\times 5$ katına çıkmasından ötürü, Bölüm 4.4.1.1'deki gibi 820 epoch'a kadar eğitilmesi mümkün olmadı. Çünkü bu veri kümesinin nispeten büyük ölçekli olmasının yanı sıra, her biri bir eğitim/test bölmesi için olmak üzere toplamda 10 kez çalıştırılması oldukça zaman almaktadır. Ancak Şekil 4.13 ve Şekil 4.14'teki eğitim raporundan görüldüğü gibi model eğitimini tamamlayarak yakınsamaktadır.



Şekil 4.13. Çoklayan-dönüştürme yaklaşımının Washington RGB-D Nesne veri kümesindeki bir bölmede yitim-iterasyon değişimini gösteren eğitim raporu.



Şekil 4.14. Çoklayan-dönüştü yaklaşımın Washington RGB-D Nesne veri kümesindeki bir bölmede doğruluk-iterasyon değişimini gösteren eğitim raporu.

2D3D Nesne veri kümesi için ise model 3000 epoch'a kadar eğitilmektedir. Her iki veri kümesi için de elde edilen sonuçlar, tek-dönüştü ve çoklu-dönüştü yaklaşım sonuçları ile beraber Çizelge 4.4'te görülmektedir. Çoklu-dönüştü yaklaşımda birden fazla giriş görüntüsü beraber ele alındığı için çizelgede * belirteci ile bu sonuç gösterilmiştir. Giriş görüntüsünün dönme matrisleri ile çoğaltılarak çoklu-dönüştü yaklaşım gibi oylamanın kullanıldığı çoklayan-dönüştü yaklaşımı, 2D3D Nesne veri kümesinde başarıyı artırırken (tek-dönüştü yaklaşıma göre), Washington RGB-D Nesne veri kümesinde başarı düşmektedir.

Çizelge 4.4. Önerilen tek-dönüştü, çoklu-dönüştü ve çoklayan-dönüştü yaklaşımlarının Washington RGB-D Nesne ve 2D3D Nesne veri kümelerinde elde ettikleri doğruluk sonuçları (%).

Yöntem	Washington RGB-D Nesne	2D3D Nesne
Bu çalışma (tek-dönüştü)	82.4 ± 2.2	90.1
Bu çalışma (çoklu-dönüştü) *	85.9 ± 2.9	94.5
Bu çalışma (çoklayan-dönüştü)	81.6 ± 1.6	93.0

4.4.4. Renk Bilgisinin Kullanıldığı Yaklaşımlar ve Deneysel Sonuçlar

Şekil 4.9 incelendiğinde, sadece derinlik bilgisinin kullanıldığı hacimsel tanımda, renkleri farklı bile olsa geometrik yapıları benzer olan nesnelerin karıştırılabildiği görülebilir. Bu bölümde, renk bilgisi kullanılarak bu tarz sınırlamaların üstesinden gelmek amaçlanmıştır. Bu kapsamda, hacimsel grid gösterimi olarak yoğunluk gridi düşünülmüştür. Bir vokseldeki renk değeri hesaplanırken, yoğunluk gridinde o vokseli ifade eden noktaların renk değerlerinin ortalaması alınmıştır. Bu amaçla, renk bilgisinin yoğunluk grid gösterimlerinde dört farklı senaryoda kodlandığı yapılar incelenmiştir: (i) piksellerin gri tonlama değerlerinin ortalamasının alındığı grid (gri-tonlamalı grid) (ii) 24-bitlik RGB renk değerlerinin 8-bitlik yapılarda kodlandığı grid (8-bit renk gridi) (iii) RGB renk değerlerinin hiperküp yapılarında ele alındığı grid (hiperküp grid) (iv) RGB değerlerinin çoklu-dönüş yaklaşımı ile ele alınması (RGB çoklu-dönüşlü yaklaşımı).

- i. **Gri-tonlamalı Grid.** Gri-tonlamalı hacimsel gösterimde bir vokselin değeri, o voksele düşen noktaların gri tonlamalı renk değerlerinin ortalamaları alınarak aşağıdaki gibi hesaplanmaktadır:

$$\tau_{ijk} = \frac{1}{N} \sum_{n=1}^N c_n \quad (4.4)$$

Buradaki N , τ_{ijk} vokselinin yoğunluk değerini, c_n ise ilgili noktanın gri tonlama değerini ifade etmektedir. Böylece her bir vokselin değeri, 0 – 255 arasında gri tonlamalı renk değeri ile ifade edilmektedir.

- ii. **8-bit Renk Gridi.** RGB görüntülerinde, her bir renk kanalı için 8 bit olmak üzere toplamda 24 bit ile ifade edilen 16,777,216 renk kodlanmaktadır. 8-bit renk gridinde, toplamda 256 renk, aşağıda her bir bitte kodlanan renk kanalına göre ifade edilmektedir.

Çizelge 4.5. 8-bitlik renk kodlanması

Bit	7	6	5	4	3	2	1	0
Renk Verisi	R	R	R	G	G	G	B	B

Bu yaklaşıma göre, önce yukarıdaki (4.4) denklemine göre bir vokseldeki ortalama RGB renk değeri bulunur. Ancak gri-tonlamalı rengi ifade eden (4.4)'deki denklemde c_n , 1×1 'lik 0 – 255 arasında değişen bir değer iken, burada hesaplanacak olan c_n , 1×3 'lük her biri 0 – 255 arasında olan RGB değerlerini

ifade eden bir dizidir. Daha sonra RGB değeri bulunan vokselin, 8-bitlik renk kodlanması Çizelge 4.5'teki renk kanallarının bit dağılımına göre aşağıdaki gibi hesaplanmaktadır.

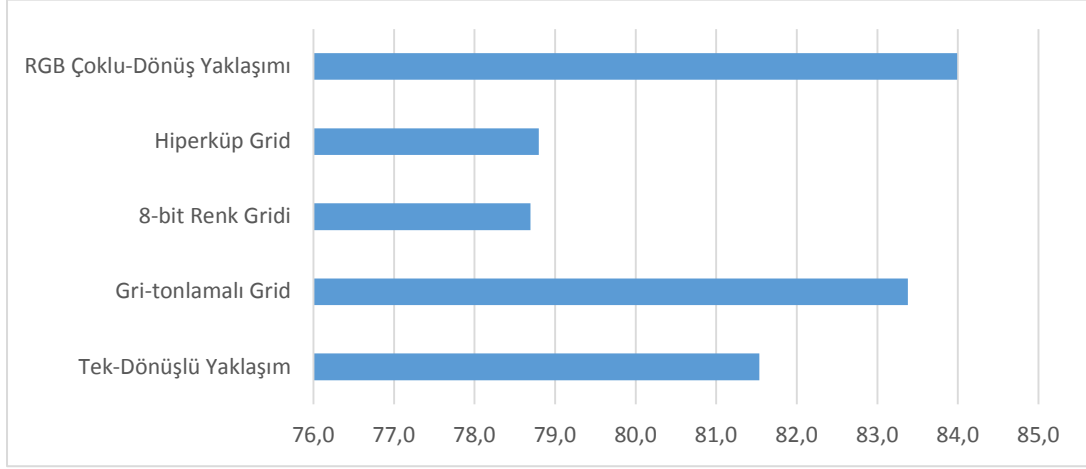
$$\tau_{ijk} = \left(\frac{R}{32}\right) \ll 5 + \left(\frac{G}{32}\right) \ll 2 + \left(\frac{B}{64}\right) \quad (4.5)$$

Böylece 3'er bit Kırmızı ve Yeşil renkleri, 2 bit de Mavi rengi, toplamda 8 bitte kodlanmaktadır.

- iii. **Hiperküp Grid.** Hiperküp grid yapısında, diğer hacimsel grid yapılarından farklı olarak, bir boyut daha eklenerek 4-boyutlu bir tensör kullanılmaktadır. Böylece $30 \times 30 \times 30$ olan hacimsel temsiller, $30 \times 30 \times 30 \times 3$ boyutlarına dönüştürülerek, 4. boyutta bir vokselin ortalama RGB değerleri tutulmaktadır. Yukarıda anlatılan gridlerde olduğu gibi, bu temsilde de yoğunluk gridi kullanılarak, (4.4) denkleminde göre bir vokselin ortalama RGB değeri hesaplanmaktadır.
- iv. **RGB Çoklu-Dönüştürme Yaklaşımı.** RGB çoklu-dönüştürme yaklaşımında, RGB kanallarının her biri birer dönme görüntüsüyümüş gibi ayrı hacimsel yapılarda temsil edilip, giriş olarak ESA modeline verilmektedir. Dolayısıyla bir giriş görüntüsü aslında 3 hacimsel giriş temsili ile ifade edilir ve Bölüm 4.3.3'te anlatıldığı gibi çoklu-dönüştürme yaklaşımı ile tanıma gerçekleştirilir.

Renk bilgisinin kullanıldığı bu 4 hacimsel yaklaşım da Washington RGB-D Nesne veri kümesi kullanılarak test edilmiştir. Veri kümesi kurulumu için, [67] çalışmasındaki kurulum ayarlamaları takip edilmiştir. Deneysel veri kümesi ile sağlanan 10 eğitim/test bölmelerinin sadece 2 tanesi kullanılmış ve bu iki bölmeden elde edilen doğruluk sonuçlarının ortalaması alınarak başarı performansları ölçülmüştür. Tüm yaklaşımlarda, Bölüm 4.3.2'de anlatılan aynı 3B ESA (Şekil 4.2) mimarisi kullanılmaktadır. Modelin eğitimi 820 epoch'a kadar devam ettirilmiştir. Renk bilgisinin kullanıldığı yaklaşımların başarı sonuçları, tek-dönüştürme yaklaşımın aynı 2 eğitim/test bölmelerinden elde ettiği ortalama başarıyla beraber Şekil 4.15'te görülmektedir. Şekilden de görüleceği üzere, ek renk bilgisine rağmen başarı performansı beklenilenin aksine pek artmamaktadır. Tek-dönüştürme yaklaşımına göre başarının arttığı senaryolar, gri-tonlamalı grid yapısı ile RGB çoklu-dönüştürme yaklaşımlarıdır. Ancak bu artış oranı da rengin sağlayacağı ek avantaj düşünüldüğünde beklentinin aşağısında kalmaktadır. Bu deneylerde, adil bir karşılaştırma yapabilmek için aynı 3B ESA modeli kullanılmıştır. Bununla beraber, renk bilgisinin

kullanıldığı yaklaşımların performansını artırmak için deneysel olarak birçok başka model araştırılmış, fakat başarıyı istenilecek düzeyde artıracak bir model bulunamamıştır.



Şekil 4.15. Renk bilgisinin kullanıldığı yaklaşımların ve tek-dönüslü yaklaşımının Washington RGB-D Nesne veri kümesindeki 2 eğitim/test bölmesindeki ortalama doğruluk başarı oranları (%).

4.5. Sonuç

Bu bölümde anlatılan çalışmalarda, 3B ESA mimarisine dayanan hacimsel bir nesne tanıma yöntemi ile derinlik verilerini iki farklı hacimsel grid ile temsil eden yaklaşımlar önerilmektedir. Kinect'ten elde edilen derinlik görüntülerinin eksik 3B model sınırlamalarına rağmen, yaygın kullanılan iki veri kümesi olan Washington RGB-D ve 2D3D Nesne veri kümelerinde, önerilen yaklaşımların etkinlikleri gösterilmektedir. Çoklu-dönüslü yaklaşımı, farklı açılardan elde edilmiş görüntü ipuçlarını birleştirerek en iyi sonuçları sağlarken, tek-dönüslü yaklaşım, diğer ilgili yöntemlerin sonuçları ile rekabet eden sonuçlar vermektedir. Çoklu-dönüslü yaklaşım, bakış açısıyla ilgili belirsizlikleri aşmak için robotik alanında yararlı olabilir. Bir robot, ortamdaki bakış açılarını değiştirebilir ve önerilen yaklaşımı kullanarak tanıma doğruluğunu artırmak için nesnelerin birden fazla görüntüsünden bilgiler derleyebilir. Ayrıca giriş görüntüsünü dönme matrisleri ile çoğaltarak, çoklu-dönüslü yaklaşım gibi ele alan bir yaklaşım önerilmiştir. Çoklayan-dönüslü diye adlandırılan bu yaklaşım, küçük veri kümesi olan 2D3D Nesne veri kümesinde başarıyı önemli bir ölçüde artırırken, daha büyük ölçekli veri kümesi olan Washington RGB-D'de aynı performansı sergileyememiştir. Sadece derinlik bilgisinin kullanıldığı yaklaşımların yanı sıra, hacimsel tanımda renk bilgilerinden de faydalanan dört farklı yaklaşım sunulmuştur. Ancak deneysel sonuçlar, beklenenin aksine rengin başarıyı genel olarak artırmadığını göstermiştir. Bu durumun sebepleri halen araştırılmaktadır. Çeşitli hacimsel nesne tanıma yaklaşımlarının önerildiği bu çalışmaların,

semantik haritalama ve tanıma görevleri dahil olmak üzere robotik görme uygulamaları için yararlı olabileceği düşünülmektedir.

Öte yandan, bu bölümde anlatılan hacimsel nesne tanıma çalışmalarında, 3B ESA modeli, kullanılan veri kümelerinde sıfırdan eğitilerek ele alınmaktadır. Derin öğrenme modellerinin içerdikleri milyonlarca parametresinin en iyi şekilde yakınsanarak öğrenilmesi, büyük ölçekli veri kümelerindeki eğitime doğrudan bağlı olmaktadır. Dolayısıyla eğitim verilerinin çokluğu, veriye aç derin öğrenme modellerinde oldukça önemli olmaktadır. Kullanılan RGB-D veri kümeleri, ImageNet veri kümesi gibi milyonlarca veri içeren RGB veri kümelerine göre oldukça küçük kalmaktadır. Bu durum RGB veri kümelerinde eğitimi gerçekleştirilen modellere göre başarıyı sınırlamaktadır. Dolayısıyla, daha büyük ölçekli veri kümesi olan ImageNet RGB veri kümesinde eğitilmiş hazır bir ESA modelini kullanarak, RGB-D nesne tanıma başarısı artırılabilir. Tezin bir sonraki kesiminde, bu amaçla ImageNet veri kümesinde eğitilmiş hazır bir ESA modeli, çoklu ÖSA yapıları ile birlikte etkin bir şekilde kullanılarak, RGB-D nesne tanıma başarısı önemli ölçüde artırılmaktadır.

5. ÖNEĞİTİMLİ BİR ESA MODELİNİ ÇOKLU ÖSA YAPILARI İLE BİRLİKTE KULLANARAK RGB-D NESNE TANIMA

5.1. Giriş

Evrişimsel sinir ağları (ESA), veri kümelerinde sıfırdan eğitilerek kullanılabilirler gibi, başka bir veri kümesinde eğitilmiş bir modeli son tam-bağlantılı katmanını kaldırarak nitelik çıkartıcı olarak hazır bir şekilde kullanmak da mümkündür. Öneğitilmiş bir ESA modelinin hazır nitelik çıkartıcı olarak kullanımı, transfer öğrenme konusu olmaktadır. Son yıllarda bu tarz öneğitilmiş ESA kullanımları yaygınlaşarak, bu modellerin sundukları kullanıma hazır, etkin nitelikleri, el yapımı nitelik sunumlarının yerini alarak birçok problemin çözümlerine uygulanmalarını sağlamıştır. Derin nitelikler, biyolojik olarak esinlenmiş değerli bilgileri hazır kullanıma sundukları için, nesne tanıma (örn. [89], [90]), tespit etme (örn. [91], [92]), semantik bölütleme (örn. [91], [93]) gibi çeşitli araştırma çalışmalarının odak noktası olmuştur. Bu çalışmalarda görülen ortak bir yaklaşım, son tam-bağlantılı katmanlardan çıkartılan niteliklerin kullanılmasıdır. Bu durumun temelindeki sebep, bu niteliklerin nesne sınıflarına ilişkin anlamsal bilgileri daha küçük boyutlarla sağlamalarıdır. Ancak, son katmanlara doğru ilerlerken, bu niteliklerin giderek seçilen veri kümesine ve göreve bağlı oldukları gözlenmiştir [94]. Öte yandan, daha erken katmanlar, eldeki görevle ilgili farklı bilgileri yakalarken, semantik bilgilere daha az duyarlı olan yerel olarak etkinleştirilen nitelikleri sağlarlar [83], [95]. Daha erken katmanların bir sorunu, onlardan çıkartılan niteliklerin yüksek boyutlu olmalarıdır. Sonuç olarak, nitelikler, genelden özele doğru ağ katmanları boyunca dönüşmektedirler ve ilişkisel ilgi bilgisi ağ genelinde farklı seviyelerde dağıtılmaktadır [83], [94]. Ancak, ağ boyunca dağıtılan bu bilginin etkin bir şekilde nasıl kullanılacağı açık değildir.

Bu bölümde anlatılan çalışmanın amacı, iki önemli anlayışı birleştirerek RGB-D nesnelerini daha doğru bir şekilde sınıflandırmak için, derin nitelik tabanlı güvenilir bir yaklaşım geliştirmektir. Bu anlayışlardan birincisi, derin nitelik çıkartıcı olarak geniş ölçekli bir veri kümesinde eğitilmiş bir ESA modelini kullanmak ve daha iyi tanıma performansı elde etmek için ağıncı farklı katmanlarından elde edilen bilgilerden yararlanmaktır. İkincisi, çıkartılan ESA niteliklerinin boyutlarını düşürmek ve bu aktivasyonları güçlü hiyerarşik nitelik temsillerinde kodlamak için, özyinelemeli sinir ağlarını (ÖSA) uygulamaktır. Eğitilmiş bir ESA modelinin ÖSA yapısı ile birleştirilmesi fikri ilk olarak RGB görüntü sınıflandırması için [96] çalışmasında sunulmuştur. Çeşitli deneyler yaptıktan sonra yazarlar, ÖSA tarafından dönüştürülmüş [97] çalışmasında

sunulan öneđitimli ađın 4. katmanından elde edilen aktivasyon ađırlıklarının, RGB görüntü sınıflandırması için gürbüz temsiller sunan daha uygun yapılar oldukları sonucuna varmışlardır. Bu çalışmadaki amaç, hem RGB hem de derinlik görüntüleri için, ađdaki farklı düzeylerden nitelik gösterimlerini, kompakt ve temsil kabiliyeti yüksek bir nitelik vektöründe toplayarak, bu fikri geliştirmektir. Ancak, [96] çalışmasından farklı olarak, önerilen yaklaşımda, nihai nitelik boyutlarını düşürmek amacıyla ESA modelinde farklı düzeylerden elde edilen aktivasyon haritalarının oldukları gibi kullanılması yerine, bu aktivasyonlar, yeniden boyutlandırılarak ÖSA yapılarına verilmektedir. Böylece, performansı düşürmeden ađaç yapılarını sabitleyerek her katman için genel bir yapı sağlanırken, farklı seviyelerden elde edilen nitelik vektörlerini birleştirerek, tanıma doğruluğunun geliştirilmesine olanak sağlanmış olur. Çoklu sabit ÖSA yapılarının öneđitimli ESA modeli ile birlikte kullanımı, nesnelerin hem semantik hem de uzamsal yapılarını korumak için farklı hiyerarşik katmanlardan nitelik geçişine olanak sağlar. Ayrıca, geniş ölçekli bir RGB veri kümesi olan ImageNet'te [23] eğitilmiş ESA modelinden derinlik alanında da bilgi transferinin sağlanabilmesi için, yüzey normalleri kullanılarak RGB bilgisi, derinlik veri alanında kodlanmaktadır. Bu amaçla, derinlik haritalarından yüzey normalleri hesaplanmakta ve hesaplanan normallerin her bir boyutu bir RGB kanalı gibi ele alınarak renklendirilmektedir. Nihai RGB-D sınıflandırma sonuçlarını elde etmek için, RGB ve derinlik alanlarından ayrı ayrı elde edilen nitelik vektörleri birleştirilmektedir. Önerilen yaklaşım, popüler Washington RGB-D Nesne [67] veri kümesinde değerlendirilmekte ve mevcut en gelişkin yöntemlerle sınıflandırma doğruluđu baz alınarak kıyaslanmaktadır. Deneysel sonuçlar, önerilen yöntemin hem nitelik boyutları hem de sınıflandırma doğruluđu açısından etkin olduğunu göstermektedir. Dolayısıyla, bu bölümde anlatılan çalışmanın katkıları aşağıdaki gibi özetlenebilir:

- RGB-D nesne kategorilerini tanımak için, ÖSA ile öneđitimli bir ESA modelini birleştiren bir ardışık düzen mimarisinde, farklı katmanlardaki bilgileri kodlayan yeni bir derin nitelik tabanlı yaklaşım sunulmaktadır.
- Öneđitimli ESA modelinden elde edilen nitelikler ve önerilen yaklaşım ile üretilen nitelikler karşılaştırmalı olarak incelenmektedir. ÖSA ile, ESA'dan elde edilen aktivasyon haritalarını, performansa zarar vermeden daha küçük boyutlarda temsil edildiđi ve daha kompakt temsilleri elde etmek için birden çok hiyerarşik seviye bilgilerini kodladıđı gösterilmektedir.

- RGB görüntülerde eğitilen bir ESA modelinden derinlik verileri için transfer öğrenmeye izin vermenin bir yolu tanımlanmaktadır. Bu amaçla, derinlik haritalarından yüzey normalleri hesaplanıp normalleştirilmektedir. Derinlik ve RGB veri türlerinin karakteristik farklarına rağmen, sonuçlar RGB görüntüleri üzerinde eğitilmiş bir ESA modelinin, bu yolla derinlik görüntülerinden etkili bir şekilde bilgiler yakalayabileceğini göstermektedir.
- Önerilen yaklaşımın, Washington RGB-D Nesne veri kümesindeki en gelişkin yöntem sonuçlarını geliştirdiğini gösteren deneysel sonuçlar sağlanmaktadır.
- ÖSA ağaç yapılarını, Bölüm 4'te anlatılan hacimsel nesne tanıma modelinde de kullanarak, sadece ESA ile elde edilen sonuçlara göre, daha iyi sonuçlar verdiğini gösteren deneyler sunulmaktadır. Böylece farklı veri modellerinde ÖSA'nın etkisi araştırılmış olmaktadır.

5.2. İlgili Çalışmalar

Derin öğrenme ağlarının, girdilerin en düşük seviyelerinden ilgili bilgileri otomatik olarak elde edebilen biyolojik olarak esinlenmiş güçlü öğrenme modellerini sunmaları ve bu bilgileri eldeki bir sorun için optimize edebilme yetenekleri, bu modellerin birçok problem çözümünde kullanımlarını yaygınlaştırmıştır. Son yıllardaki çalışmalar, büyük ölçekli bir veri kümesinde eğitilmiş bir ESA modelinin, diğer görsel tanıma görevleri için de genelleyici iyi temsiller oluşturmak için etkili bir şekilde kullanılabilirliğini göstermiştir [89], [94], [98], [99]. Gupta ve diğerleri [100] büyük ölçekli RGB veri kümelerinde önceden eğitilmiş bir ESA modelini kullanmak için giriş görüntülerinin kamera parametrelerini kullanarak derinlik bilgilerini üç kanalda kodlamaktadırlar ve RGB-D nesne tespiti problemine odaklanmaktadır. Schwarz ve diğerleri [90], Krizhevsky ve diğerlerinin [22] öneğitilmiş ESA modelinin son tam-bağlantılı katmanları olan $fc7$ ve $fc8$ aktivasyonlarını kullanarak RGB-D nesne tanıma ve poz tahmini için bir yaklaşım sunarlar. Eitel ve diğerleri tarafından önerilen [101] çalışmasında farklı bir yaklaşım önerilmektedir. RGB ve derinlik veri türlerinin her biri için olmak üzere iki akışlı bir ESA modelini, son olarak bir geç füzyon ağıyla birleştiren bir yaklaşım kullanılmaktadır. Her iki akış da ImageNet'te öneğitilmiş bir model ağırlıkları ile ilklendirip, nihai sınıflandırma için ince-ayarlıma (*fine-tuning*) yapılmaktadır. Yakın zamandaki [83] çalışmasında ise ESA modelinin tüm katmanlarındaki aktivasyonları kodlamak için farklı düzeylerde bir uzamsal piramit havuzu stratejisi, nihai nitelik temsillerinin birleştirilmesi öncesinde kullanılmaktadır. Bu çalışmadaki farklı düzeylerden elde edilen bilgileri birleştirme

düşüncesi, bu çalışmanın bir esin kaynağını oluşturmaktadır. Asif ve diğerleri [102], öneğitimli VGGnet [103] olarak adlandırılan modelden $fc7$ niteliklerini çıkartarak, nesnelere görünüşlerini ve yapısal bilgilerini kodlayan bir yaklaşım sunarlar.

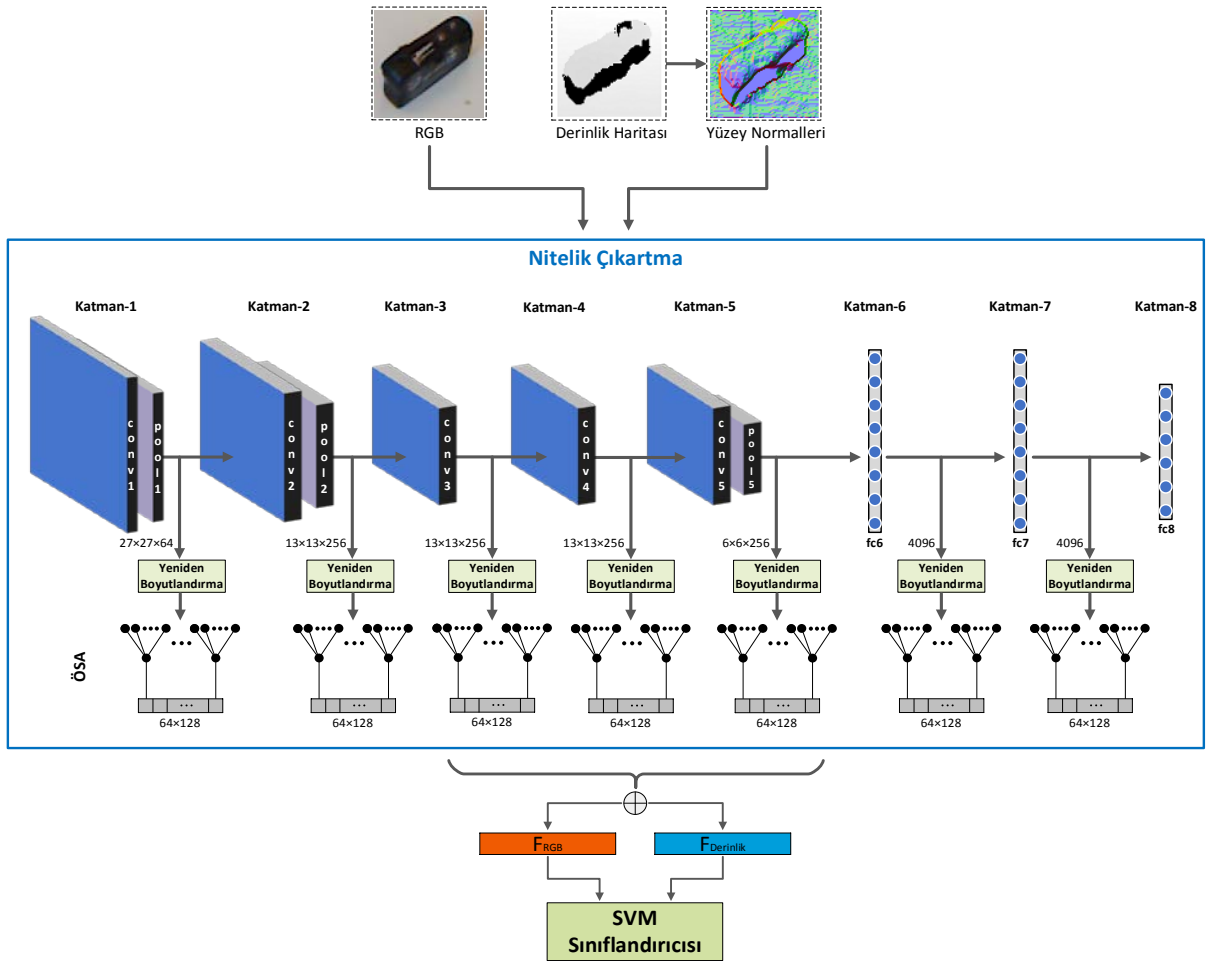
Yukarıda anlatılan derin öğrenme niteliklerine dayalı yöntemlerden başka yöntemler de geliştirilmiştir. Evrişimsel fisher çekirdekleri (*convolutional fisher kernels - CFK*) [104], ESA ile fisher çekirdeklerini entegre eden bir RGB-D nesne tanıma yöntemini sunmaktadır. Tanıma doğruluğu performansı açısından başarılı olmasına rağmen, sınıflandırma için kullanılan nihai nitelik vektörünün çok büyük boyutlu olması bu yöntemin dezavantajı olarak göze çarpmaktadır. Zia ve diğerleri [105], derinlik görüntülerindeki 3B uzamsal bilgilerden tam olarak yararlanabilmek için 3B ESA'ı ve RGB bilgilerinden yararlanmak için öneğitimli VGGnet [103] modelini kullanan bir yöntem önermektedirler. Ayrıca, öneğitimli 2B ESA ile ilklendirilmiş ve daha sonra ince-ayarlanmış bir hibrit 2B/3B ESA modeli önermektedirler. Son olarak, bu hibrit model yapısından elde edilen nitelikleri, yalnızca derinlik ve yalnızca RGB görüntüleri için kullandıkları modellerden elde edilen niteliklerle birleştirerek, oluşturdukları vektör ile nihai tanıma performansını ölçmektedirler.

Özyinelemeli sinir ağları (ÖSA) [50], [52], yapılandırılmış bilgilerden dağıtılmış temsilleri özyinelemeli ağ yapısına dönüştürülmüş çizgeler vasıtasıyla öğrenmektedirler ve çeşitli araştırma amaçlı diğer mimarilerle birlikte kullanılmışlardır [52], [56], [61], [64], [106], [107]. CNN-RNN [56] çalışmasında, Socher ve diğerleri, RGB ve derinlik niteliklerini ayrı aşamalarda öğrenen ve sonraki aşamada bu nitelikleri birleştirerek nihai RGB-D tanıma gerçekleştiren, ESA ile ÖSA modellerinin işbirliğini kullanan bir RGB-D nesne tanıma yöntemini ilk kez tanıtmaktadırlar. Daha sonra bu fikir, Bui ve diğerleri [96] tarafından tek bir katmandan oluşan ESA yapısını, öneğitimli bir ESA modeli ile değiştirmek suretiyle, RGB görüntülerde tanıma için genişletilmiştir. AlexNet-RNN olarak adlandırılan bu yöntemin başarısı, öneğitimli bir ESA modelden elde edilen niteliklerin özyinelemeli ağ yapıları ile dönüştürülmesinin RGB nesne tanıma sınıflandırma doğruluğunu önemli ölçüde artırdığını göstermektedir. Bu çalışmada, [96]'da sunulan fikir, RGB-D nesne tanıma için yeni bir yapı ile uyarlanmakta ve [83], [95] çalışmalarındaki fikirleri takiben çoklu katmanlardan çıkartılan bilgilerden yararlanılmaktadır. Bu bakımdan önerilen bu yeni yaklaşım, nesnelere gürbüz temsillerini sunar. Yaygın kullanılan Washington RGB-D Nesne [67] veri kümesinde yapılan deneysel değerlendirmeler, önerilen yaklaşımın nitelik

boyutlarını önemli oranda düşürerek doğruluk performansını artırdığını göstererek, RGB-D nesne tanıma için etkinliğini ortaya koymaktadır.

5.3. Önerilen Yöntem

Bu çalışmada, RGB-D verilerinde nesne kategorilerini tanımak üzere öneğitilmiş bir ESA modelinin ÖSA yapıları ile birlikte kullanılmasının etkinliği araştırılmaktadır. Özellikle nesne tanımda yaygın olarak kullanılan (örn. [83], [96]) VGG-f [97] olarak anlandırılan öneğitilmiş ESA modeli kullanılmaktadır. Derinlik verileri için ImageNet [23] gibi büyük ölçekli RGB veri kümeleri üzerinde önceden eğitilmiş ESA modellerinin gücünden yararlanmak için, her bir pikselde üç renk kanalını kodlamak üzere derinlik görüntüleri bir önışleme tabi tutulmaktadır. Bu amaçla ilk önce, derinlik haritalarından, her bir boyutu bir renk kanalını temsil etmek üzere üç boyutlu yüzey normalleri hesaplanmaktadır. Daha sonra, her bir kanal için değerleri 0 – 255 renk aralığına eşlemek üzere veriler normalleştirilmektedir.



Şekil 5.1. Önerilen yaklaşımın genel görünümü. Giriş olarak RGB görüntüleri ve renklendirilmiş yüzey normalleri görüntüleri alınmaktadır. Farklı katmanlardan ham nitelikler çıkartmak üzere öneğitilmiş bir ESA modeli [97] kullanılmaktadır. Çoklu ÖSA,

sabit ağaç yapılarında daha yüksek seviyeli temsiller elde edilmek üzere kullanılmaktadır. Farklı katmanlardan elde edilen temsiller, RGB ve derinlik veri alanlarında nihai nitelik vektörlerini oluşturmak üzere birleştirilmektedir ve doğrusal bir SVM sınıflandırıcısına verilmektedir.

Önerilen yaklaşım, genel yapısı ile beraber Şekil 5.1’de gösterilmektedir. Önerilen yaklaşım, iki aşamalı bir hiyerarşik derin nitelik çıkartma işlem grubunu içermektedir. İlk adımda, aktivasyon haritaları, ötelemede değişmez yararlı nitelikleri (*translational invariant features*) yakalamak üzere öneğitimli bir ESA modelinden farklı seviyelerde çıkartılmaktadır. Daha sonra, bu aktivasyon haritaları, nihai boyutlarının düşürülmesi için yeniden boyutlandırılarak, çoklu sabit ÖSA ağaç yapılarına, görüntülerin hiyerarşik yüksek seviyeli niteliklerine eşlenmek üzere verilmektedir. Bu amaçla, Bui ve diğerleri tarafından önerilen çalışma [96], uyarlanarak kapsamı genişletilmektedir. [96] çalışmasında yazarlar, öneğitimli bir ESA modelini ÖSA ile beraber, bir RGB-D veri kümesinde yalnızca RGB görüntülerinde kullanmaktadırlar. Yaklaşımlarında, ESA modelinin bir katmanından elde ettikleri aktivasyon haritalarını olduğu gibi özyinelemeli ağ yapısına giriş olarak vermektedirler. Buradaki ayarlamaların aksine, bu çalışmada, birbirlerini tamamlayıcı farklı nitelik örüntülerini elde etmek için hem RGB hem de derinlik görüntülerinde birçok seviyeden sağlanan niteliklerin birleştirilmesi istenmektedir. Bu nedenle, temel alınan yapı birkaç açıdan değiştirilerek ele alınmaktadır. İlk olarak, ESA modelinden elde edilen aktivasyon haritaları, yeniden boyutlandırılarak ÖSA ile üretilen nitelik vektörünün yüksek boyutluluğu ile başa çıkılmaktadır. Böylece, bu durum daha fazla sınıflandırma performansı elde etmek için farklı katmanlardaki bilgilerden yararlanma olanağı verir. Bu şekilde, birden fazla katman, her nesne sınıfı için kompakt ve temsil kabiliyeti yüksek bir nitelik vektörü sağlar. İkinci olarak, derinlik haritalarından yüzey normaleri hesaplanarak ve RGB veri türüne benzer renk bilgisi kodlanarak, büyük ölçekli RGB veri kümesi olan ImageNet veri kümesinden derinlik veri türü için transfer öğrenme ile bilgi aktarımı sağlanmaktadır. Son olarak, yüksek doğruluk derecesinde RGB-D nesne kategorilerini tanımak üzere, RGB ve derinlik veri türleri için ayrı ayrı elde edilen nitelik vektörleri nihai RGB-D nitelik vektöründe birleştirilmektedir.

Öneğitimli VGG-f modeli, ince-ayarlı yapılmadan nitelik çıkartıcı bir model olarak önerilen yaklaşımın temelinde kullanılmaktadır. Dolayısıyla, bu nitelik çıkartma aşaması, herhangi bir eğitim gerektirmeksizin hızlı bir şekilde çalışır. Kullanılan model ağı, birbirlerini izleyen 5 evrişim katmanı (her biri; evrişim, havuzlama ve yerel kontrast normalleştirme işlemleri de dahil olmak üzere alt modüllere sahip olabilir) ile bunları

izleyen 3 tam-bağlantılı katmandan oluşur ve ImageNet veri kümesi üzerinde bir dağılım oluşturur. Katmanlardan elde edilen aktivasyon haritalarının boyutları sırasıyla $27 \times 27 \times 64$, $13 \times 13 \times 256$, $13 \times 13 \times 256$, $13 \times 13 \times 256$, $6 \times 6 \times 256$, 4096, 4096 ve 1000'dir. Son tam-bağlantılı katman çıktısı, ImageNet'in 1000 sınıfı üzerindeki nitelik temsillerini ifade eder. Diğer katmanlar için, filtre bankası sayısı 64'e sabitlenerek ve ÖSA ağaç yapılarının girişlerine uygun bir şekilde, aktivasyonlar, yeniden boyutlandırılmaktadır. Böylece, örneğin, tam-bağlantılı katmanların çıktıları $8 \times 8 \times 64$ formuna dönüştürülürken, aynı boyutlarda çıktı üreten evrişim katmanlarının yeni boyutları $26 \times 26 \times 64$ şeklinde olur. Böylece, bu yeni yapı formları ile genel bir kullanım kolaylığı sağlanırken, performanstan ödün vermeden ÖSA tarafından üretilen nitelik vektörlerinin boyutları küçültülmüş olur. Kullanılan öneğitilmiş ESA modeline ilişkin daha detaylı bilgi almak için, okuyucular [97] çalışmasına başvurabilirler.

Öneğitilmiş ESA modeli ile giriş görüntülerinden, ileri yayılım (*forward propagation*) ile aktivasyon haritaları elde edildikten sonra, bunlardan kompakt genel nitelik temsillerini elde etmek üzere, girişleri ESA çıktıları olan ÖSA yapıları kullanılmaktadır. ÖSA yapıları iyi çalışılmış modeller olup (örn. [50], [52], [56] gibi) aynı işlemleri bir ağaç yapısı içerisinde özyinelemeli bir şekilde uygulayarak, daha yüksek seviyeli temsilleri elde edebilirler. Her bir seviyede komşu vektörler bir üst vektörde bağlı ağırlıklar kullanımı ile birleştirilerek, nihayetinde $X \in \mathbb{R}^{K \times r \times r}$ olan girişlerin, çoklu katmanlardan ilerleyerek daha düşük boyutlu bir uzaydaki $p \in \mathbb{R}^K$ ata vektöre eşlenmesi amaçlanmaktadır. Daha sonra, ata vektör doğrusal olmayan bir ezme (*squash*) işlevinden geçirilir. Bu çalışmada, ÖSA'nın önerildiği çalışmada kullanılan tanh, orijinal çalışmayı korumak için kullanılmaktadır. Ancak yeterli doğrusalsızlığı sağlayan herhangi bir ezme işlevi kullanılabilir (örn. Hiperbolik tanjant sigmoid veya elliot sigmoid fonksiyonları benzer sonuçlar üretmektedir). Tek bir ÖSA yapısı K -boyutlu bir nitelik vektörünü üretir. Buradaki K girdi olarak verilen verinin büyüklüğünü, diğer bir ifadeyle filtre bankası boyutunu ifade etmektedir. Bu çalışmada, çoklu rastgele ilklendirilmiş N adet ÖSA yapıları kullanılmaktadır. Bu nedenle, nihai vektör olarak bu işlemler sonucunda $(N \times K)$ -boyutlu temsiller elde edilmektedir.

ÖSA yapılarının bu süreçteki rolü iki yönlüdür. İlk olarak, nitelik boyutlarını düşürmekte ve sınıflandırma performansını en üst düzeye çıkartmaktadır. Böylece, birden çok katmandan elde edilen bilgileri etkili bir şekilde nihai sınıflandırmada kullanılmak üzere aktarmayı sağlar. İkincisi, sezgisel olarak çocuk düğümlerinin semantik içeriği, bu yapılar

boyunca ata düğümlerde özyinelemeli bir şekilde toplanmaktadır. Bu şekilde, ortaya çıkan bilgi nesne görüntüsünün tamamına ilişkin bağlamı temsil eder. Ayrıca, kullanılan ÖSA yapıları geriyayılım kullanılmadan rastgele iklendirilmiş ağırlık kullanımlarına dayanmaktadır. ESA'daki alıcı alan kullanımından farklı olarak, ÖSA yapılarında birbirleri ile örtüşmeyen (*non-overlapping*) alıcı alanlar kullanılmaktadır. Ek olarak bu çalışmada özellikle, tek seviyeli sadece bir ata vektörden oluşan ÖSA yapıları kullanılmaktadır. Böylece, bu aşamada nitelik temsillerinin elde edilmeleri son derece hızlı olmaktadır.

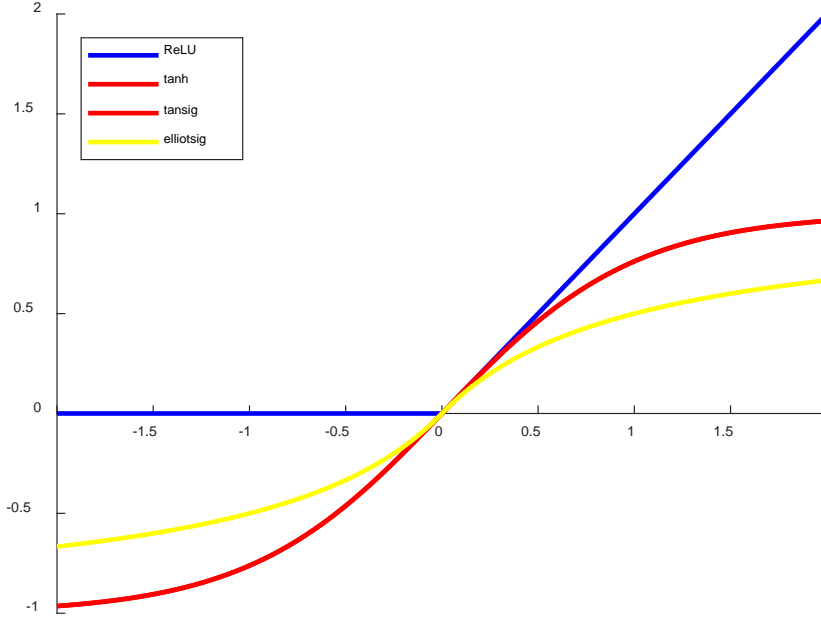
5.4. Deneysel Değerlendirmeler

Önerilen yöntem, Washington RGB-D Nesne [67] veri kümesi kullanılarak değerlendirilmektedir. Deneyler, [67]'de sağlanan 10 eğitim/test bölmesi kullanılarak gerçekleştirilmektedir. Her bir bölmede, yaklaşık 35,000 eğitim görüntüsü ve 7,000 test görüntüsü mevcuttur. Veri kümesinin kullanımına uygun olarak, her kategoriden bir alt örneklem test için ayrılırken, geri kalanlar eğitim için kullanılmaktadır. Giriş görüntülerinin ele alındığı VGG-f modeline uygun olması için görüntüler 224×224 piksel boyutlarına yeniden boyutlandırılmaktadır. Kullanılan veri kümesinde ayrıca nesne bölütlerine ilişkin maskeler de sağlanmaktadır. Ancak görüntülerin arkaplanları karmaşık olmayıp, tüm görüntüler için sabit ve basit bir yapıdadır. Bu yüzden arkaplandan kurtulmak için fazladan bir önışlem adımı uygulanmamaktadır. Ayrıca önerilen derin öğrenme yaklaşımı, arkaplanı kolaylıkla ele alabilmektedir. Bu bölümde ilk önce deneysel sonuçlarla birlikte model analizi yapılmaktadır. Daha sonra, önerilen yaklaşımın kategori tanıma performansı, kullanılan veri kümesindeki diğer ilgili modern yöntem sonuçları ile karşılaştırılmaktadır. Önerilen yöntemde, açık kaynak kodlu MatConvNet çatısı [108] ve bununla sağlanan öneğitimli VGG-f ESA modeli kullanılmaktadır. Elde edilen nitelik temsilleri doğrusal bir SVM sınıflandırıcısı (Liblinear [109]) kullanılarak sınıflandırılmaktadır.

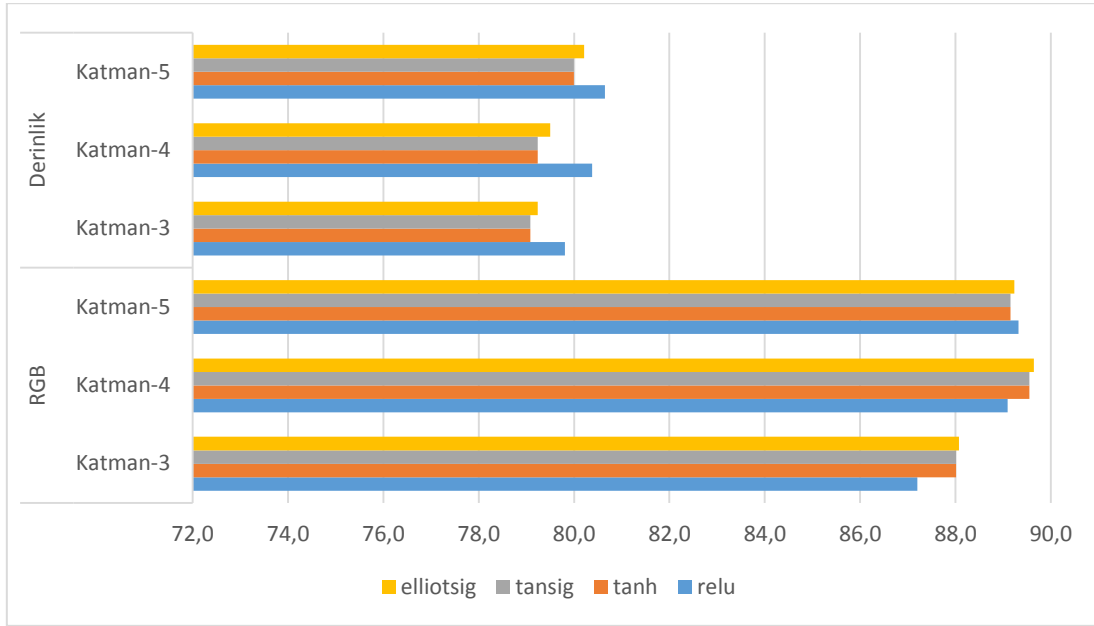
5.4.1. Model Analizi

Önerilen yaklaşım, çeşitli model farklılıklarına göre analiz edilmektedir. İlk olarak, çeşitli ezme işlevleri kullanımlarının doğruluk perfomansına olan etkileri incelenmektedir. Bu amaçla, *ReLU*, *tanh*, *tansig* ve *elliotsig* fonksiyonları olmak üzere 4 farklı doğrusal olmayan işlev kullanılmaktadır. Şekil 5.2, bu fonksiyonların verileri ele alışlarına ilişkin grafikleri göstermektedir. Deneylerde, geçerli bir karşılaştırma yapabilmek amacıyla ÖSA için kullanılacak ağırlık değerlerinin eşit olmasına dikkat edilerek, aynı rastgele ağırlık değerleri kullanılmıştır. Şekil 5.3'te sonuçlar gösterilmektedir. Sonuçların genel olarak

birbirlerine yakın oldukları görülmektedir. Bununla birlikte, ReLU ve diğer doğrusal olmayan işlevler arasında küçük bir fark vardır. ReLU fonksiyonu derinlik verileri için daha iyi sonuçlar verirken, diğerleri RGB verileri için daha iyi bir başarı elde ederler. Fark göz ardı edilebilir bir oranda olduğundan, bu çalışmada orijinal ÖSA çalışmasına [56] uygun olarak doğrusal olmayan *tanh* işlevi kullanılmaktadır.

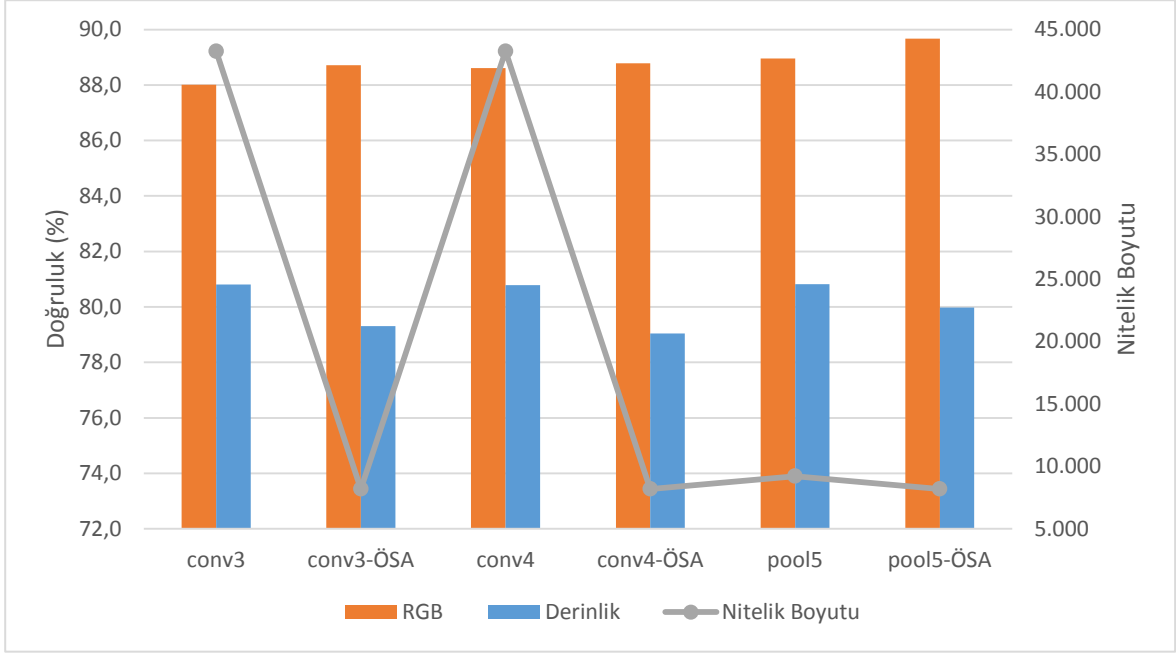


Şekil 5.2. Özyinelemeli sinir ağlarında kullanılan doğrusal olmayan fonksiyonlar. *tanh* ve *tansig* işlevleri çok küçük numerik farklarla sonuçlar üretmektedir. Bu yüzden şekilde renkler çakışmaktadır. Ancak işlem hızı olarak bu iki fonksiyon arasında farklar olabilmektedir.



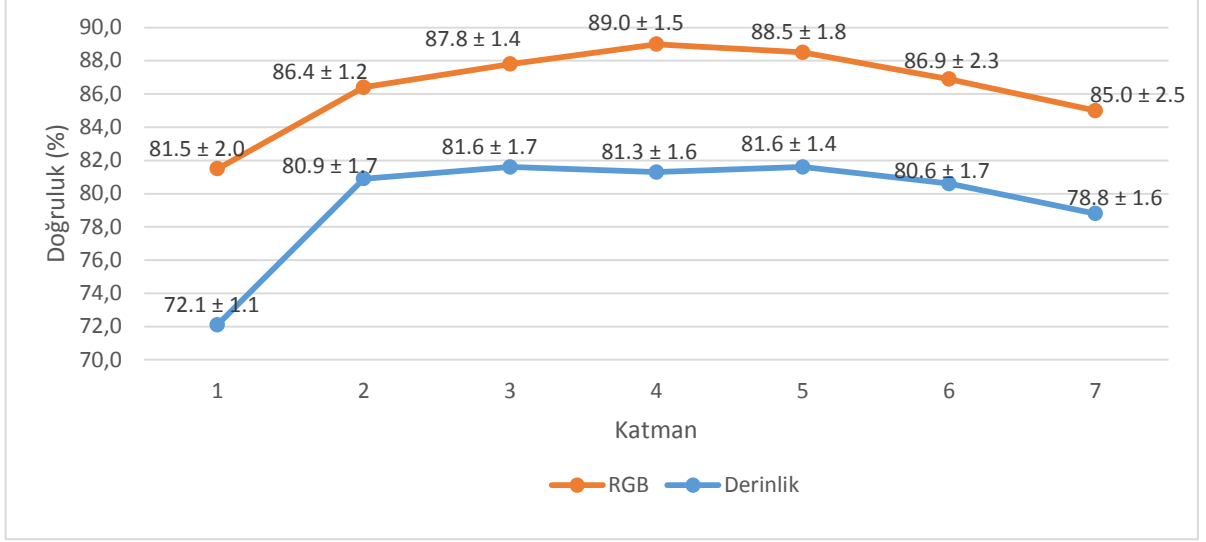
Şekil 5.3. ÖSA için farklı ezme işlevlerinin sınıflandırma doğruluğu açısından etkileri (%).

Daha sonra ÖSA'nın orta katmanlardaki (conv3, conv4 ve pool5 katmanları) etkisi değerlendirilmektedir. Şekil 5.4'te görüleceği üzere, ÖSA, ESA'dan elde edilen nitelik boyutlarını önemli ölçüde azaltarak (özellikle 3. ve 4. katmanlar için nitelik boyutları $\times 5$ kattan fazla düşürülmektedir) sınıflandırma performansını geliştirmektedir. Ancak doğruluk performanslarına bakıldığında, RGB için başarı artarken derinlik verileri için küçük bir azalma görülmektedir ($\sim 1\%$). Bununla birlikte ÖSA ile elde edilen kompakt nitelik temsilleri, sınıflandırmada hesaplama maliyetini önemli ölçüde düşürdüğü ve daha üstün performans elde etmek için çoklu katmanların çıktılarından yararlanmaya olanak verdiği için tercih edilmektedir. Hem ÖSA etkisinin değerlendirildiği bu deneylerde hem de yukarıda anlatılan ÖSA için farklı ezme işlevlerinin etkilerinin incelendiği deneylerde, veri kümesi ile sağlanan 10 eğitim/test bölmelerinden bir tanesi geliştirme bölgesi olarak kullanılmıştır. Bundan sonra anlatılacak olan geri kalan deneylerimizde 10 eğitim/test bölmelerinin tamamı, veri kümesi ayarlamalarına uygun olarak kullanılıp, ortalama sonuçlar standart sapma değerleri ile beraber verilmektedir.



Şekil 5.4. ÖSA'nın doğruluk performansı ve nitelik boyutları açısından öneğitimli ESA'dan elde edilen orta katman ham niteliklerindeki etkileri.

Deneşlerde, özellikle orta katmanlara odaklanılmaktadır. Bunun nedeni ise orta katmanlardan elde edilen çıktının en uygun temsilleri oluşturacaklarına dair sezgisel varsayımdır. Çünkü derin ağılardan elde edilen niteliklerin genelden özele doğru ağ boyunca dönüştükleri, önceki çalışmalarda gösterilmiştir [89], [110]. Erken katmanlar köşeler ve kenarlar gibi düşük düzeydeki ham niteliklere yanıt verirlerken, geç katmanlar eğitimin gerçekleştirildiği veri kümesindeki nesne sınıflarına özgü daha genel niteliklere yanıt vermektedirler. Dolayısıyla ağın orta katmanları, en uygun temsilleri sunarlar. Şekil 5.5, önerilen yöntem ile her bir katmandan elde edilen ortalama başarı oranlarını standart sapma değerleri ile birlikte göstermektedir. Burada sunulan grafik, anılan sezgisel varsayımı, başlangıcındaki net yükseliş ve sonundaki net azalış eğilimleri ile doğrulamaktadır.



Şekil 5.5. Önerilen yaklaşımın tekil katmanlar için elde ettiği doğruluk performansı.

Şimdi, bu orta düzeydeki temsillerin çeşitli kombinasyonları üzerindeki doğruluk performansının deneysel analizine geçiyoruz. Çizelge 5.1'de, farklı seviyelerdeki nitelik temsillerini birleştirmenin, tanıma doğruluğunu önemli ölçüde artırdığını gösteren sonuçlar sunulmaktadır. RGB için 4. ve 5. seviyelerden alınan nitelik kombinasyonu en iyi doğruluğu verirken, derinlik verileri için 3., 4. ve 5. seviye temsilleri birlikte en iyi sonucu vermektedir.

Çizelge 5.1. Orta seviye katmanlarının birleştirilmeleri ile elde edilen farklı kombinasyonlar için RGB ve derinlik verilerinde elde edilen doğruluk performansları (%).

	RGB	Derinlik	Nitelik Boyutu
Katman3 + Katman4	89.0 ± 1.4	83.0 ± 1.7	16,384
Katman3 + Katman5	89.4 ± 1.5	83.5 ± 1.7	16,384
Katman4 + Katman5	89.9 ± 1.6	83.4 ± 1.7	16,384
Katman3 + Katman4 + Katman5	89.8 ± 1.5	84.0 ± 1.8	24,576

Son olarak, nihai RGB-D doğruluk performansını değerlendirmek için, RGB ve derinlik nitelikleri birleştirilerek deneyler yapılmaktadır. Bu amaçla, ilk olarak, RGB ve derinlik verileri için ayrı ayrı en iyi sonuçları sağlayan tekil katmanların çıktı temsilleri birleştirilmektedir. Şekil 5.5'ten de görüleceği üzere, tekil katman performansları baz alındığında RGB için 4. katman, derinlik için 5. katman en iyi sonucu vermektedir. Daha sonra, Çizelge 5.1'deki sonuçlar dikkate alınarak iki katman birleşimi ile elde edilen sonuçlara göre, hem RGB ve hem de derinlik için en uygun sonuç veren kombinasyon

birlikte değerlendirilmektedir. Çizelge 5.2’de doğruluk sonuçları, standart sapma değerleri ile birlikte sunulmaktadır. Ayrıca nitelik boyutları son sütunda raporlanmaktadır. Daha fazla katman birleşimi ile elde edilen niteliklerin birlikte düşünülmesi, nitelik boyutunu daha da artırabilmektedir. Bu durum, daha büyük nitelik uzayında sınırlı hesaplama kaynaklarıyla sınıflandırma işlemini zorlaştırmaktadır. Bu yüzden, önerilen yaklaşımın sağladığı yüksek tanıma doğruluğu avantajının, daha büyük boyutlu nitelik uzayı ile kaybolmaması için daha fazla katman çıktısı birlikte düşünülmemektedir.

Çizelge 5.2. Nihai RGB-D tanıma doğruluk performansı için, RGB ve derinlik verilerinin birlikte değerlendirilmeleri ile elde edilen sonuçlar. RGB₄, RGB verileri için 4. katmandan elde edilen çıktılar değerlendirildiğini ifade ederken, RGB₍₄₊₅₎, RGB verilerinde 4. ve 5. katman çıktılarının birlikte değerlendirildiği kombinasyonu ifade etmektedir.

	Doğruluk (%)	Nitelik Boyutu
RGB ₄ + Derinlik ₅	92.0 ± 1.3	16,384
RGB ₍₄₊₅₎ + Derinlik ₍₄₊₅₎	92.5 ± 1.2	32,768

Deneyle yapılırken, görüntülerin normalleştirilerek işlenmeleri ile normalleştirilmeden işlenmeleri durumlarında elde edilen doğruluk performansları arasında küçük bir fark olduğu gözlemlenmiştir. Görüntüler, normalleştirilerek ele alındıklarında, RGB için elde edilen en yüksek doğruluk oranı 0.2% oranında azalırken, derinlik için 0.3% oranında artmaktadır. Her ne kadar bu fark çok küçük olsa da gözlenen bu durumu dikkate alarak, bütün deneylerde, RGB görüntüleri normalleştirme yapılmadan ele alınmaktadır. Derinlik görüntüleri için ise ImageNet ortalama görüntüsü dikkate alınarak görüntü normalleştirilmesi uygulanmaktadır.

5.4.2. Karşılaştırmalı Sonuçlar

Çizelge 5.3’te, önerilen yaklaşım ile elde edilen en iyi sonuçlar, Washington RGB-D Nesne [67] veri kümesindeki diğer ilgili en gelişkin yöntemlerin sonuçları ile karşılaştırılmaktadır. Diğer yöntemlerin sonuçları orijinal makalelerde rapor edilen sonuçlardır. Çizelgede görüldüğü gibi önerilen yaklaşım, RGB ve RGB-D verileri için en iyi sonuçları verirken, derinlik verileri için ise oldukça rekabetçi sonuçlar üretmektedir. Önerilen yaklaşımın RGB sonucuna en yakın performansı veren AlexNet-RNN [96]

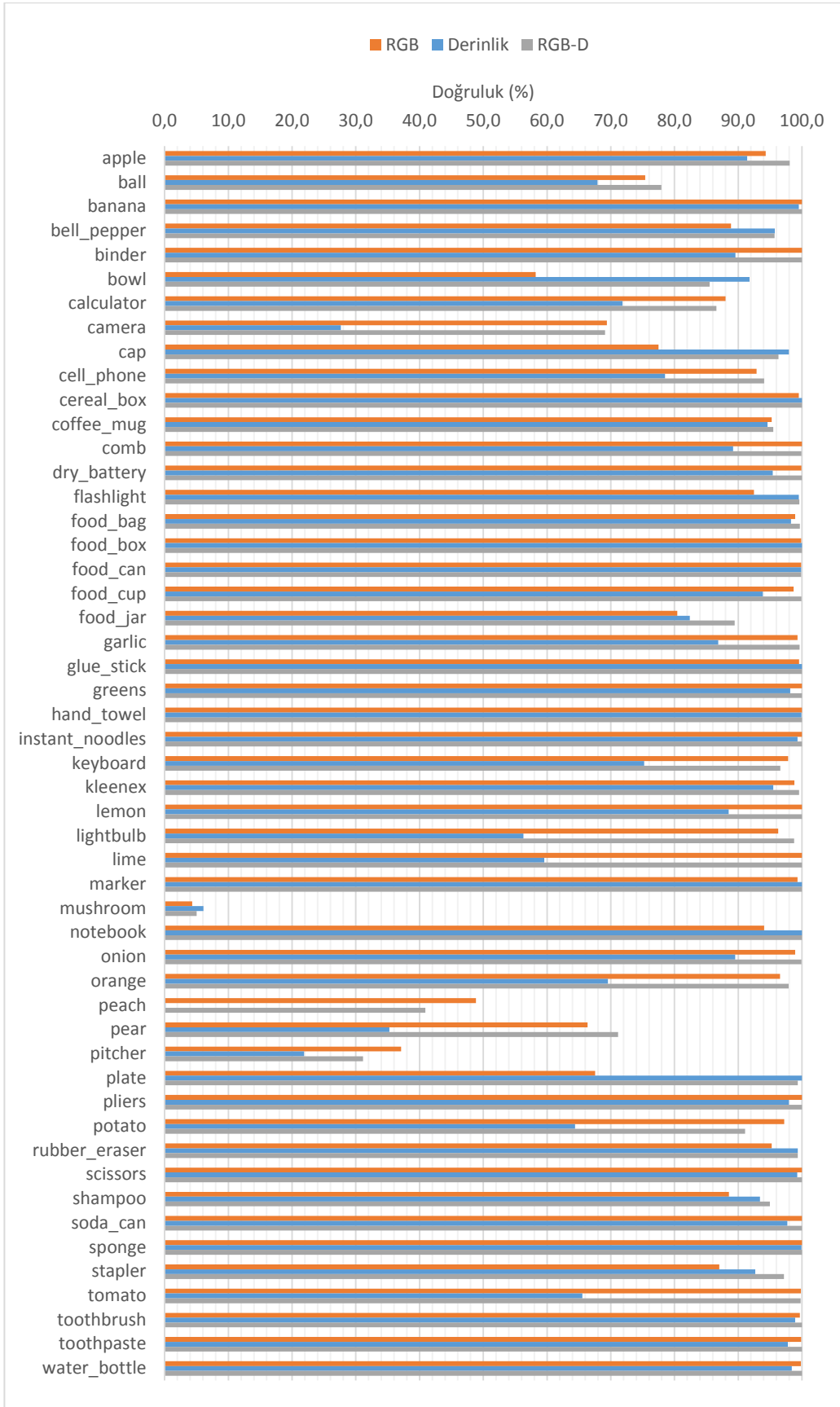
çalışmasındaki nitelik boyutu, bu çalışmada üretilen nitelik boyutunun iki katı büyüklüğündedir. Derinlik verileri için ise bu çalışmada elde edilen tanıma başarısı ilgili diğer çalışmaların çoğunluğunu geçerek, sadece Hypercube [83] ve CFK [104] çalışmalarına geçilmiştir. Hypercube [83] çalışmasında, önerilen saklı nokta bulutu renklendirilmesinde, RGB görüntülerinden yararlanılmaktadır. Dolayısıyla çizelgedeki diğer yöntemlerin aksine, bu çalışmada rapor edilen derinlik sonucu saf derinlik bilgisine dayanmamaktadır. CFK [104] yönteminde ise temel bileşenler analizi [111] ile nitelik boyutu düşürülmesine rağmen, kullanılan nitelik boyutu bu çalışmada kullanılan derinlik nitelik boyutunun yaklaşık $\times 64$ katı olan 1,568,000 büyüklüğündedir. Ayrıca temel bileşenler analizi ile veri dağılımının incelenerek büyük ölçekli uzaydan (CFK çalışmasında boyut düşürülmeden önce veri 2,508,800 boyutludur), daha düşük ölçekli bir uzaya verinin yansıtılması maliyeti göz önüne alındığında, derinlik verileri için de bu çalışmada önerilen yaklaşımın açık bir avantaj sağladığı görülmektedir. Öte yandan, RGB'ye göre derinlik veri türü için elde edilen nispeten daha düşük başarının bir nedeni de, yaklaşımın temelinde nitelik çıkartıcı olarak RGB veri kümesi olan ImageNet üzerinde eğitilmiş bir ESA modeli kullanımı olabilir. Her ne kadar derinlik verileri için yüzey normalleri kullanılarak RGB'ye benzer bir renklendirme yapılmış olsa da, RGB ve derinlik veri türlerinin farklı doğalarından dolayı, bu yöntem yüzde yüz verimlilik sağlamayabilir. Sonuç olarak, önerilen yöntem, nitelik çıkartma aşaması için herhangi bir eğitim gerektirmeksizin, hızlı ve etkili bir şekilde, nesnelere temsil kabiliyeti yüksek, ayırt edici derin nitelikler ile temsil etmektedir ve üstün doğruluk performansını düşük veri boyutları ile beraber sağlamaktadır.

Çizelge 5.3. Önerilen yaklaşımın, Washington RGB-D Nesne veri kümesindeki ilgili diğer yöntemlerle doğruluk oranı karşılaştırması (%).

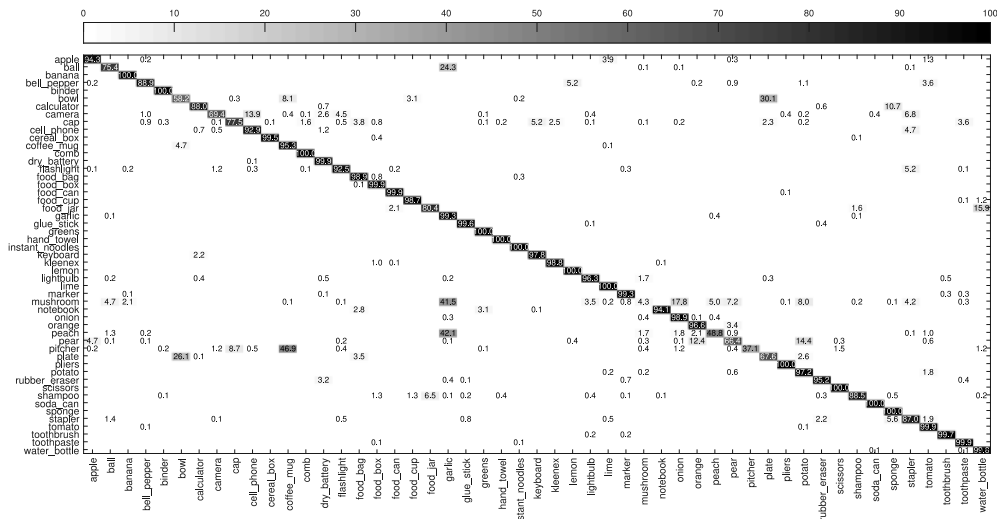
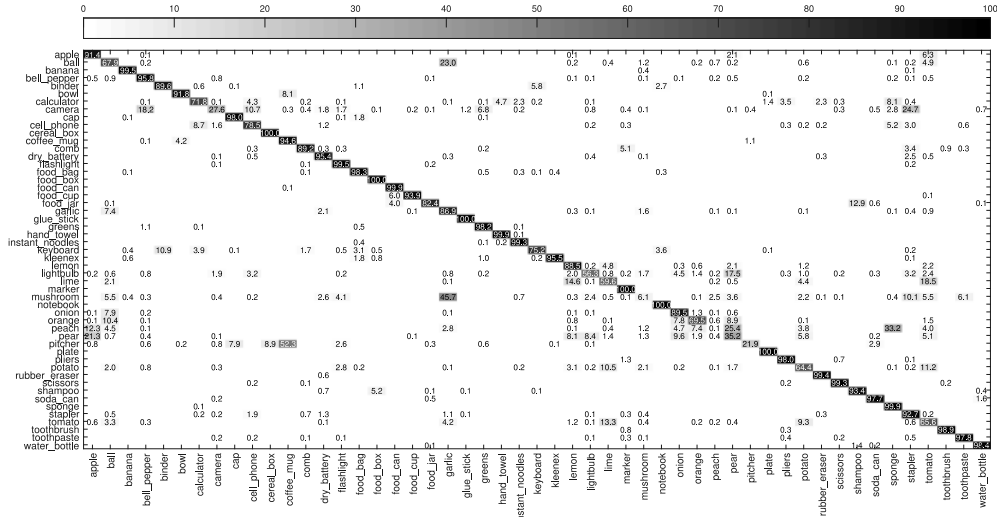
Yöntem	RGB	Derinlik	RGB-D
Kernel SVM [67]	74.5 ± 3.1	64.7 ± 2.2	83.9 ± 3.5
HKDES [81]	76.1 ± 2.2	75.7 ± 2.6	84.1 ± 2.2
KDES [72]	77.7 ± 1.9	78.8 ± 2.7	86.2 ± 2.1
CKM [60]	-	-	86.4 ± 2.3
CNN-RNN [56]	80.8 ± 4.2	78.9 ± 3.8	86.8 ± 3.3
Subset-RNN [61]	82.8 ± 3.4	81.8 ± 2.6	88.5 ± 3.1
CNN Features [90]	83.1 ± 2.0	-	89.4 ± 1.3
MM-LRF-ELM [112]	84.3 ± 3.2	82.9 ± 2.5	89.6 ± 2.5
CNN-SPM-RNN [64]	85.2 ± 1.2	83.6 ± 2.3	90.7 ± 1.1
Hypercube [83]	87.6 ± 2.2	85.0 ± 2.1	91.1 ± 1.4
CFK [104]	86.8 ± 2.7	85.8 ± 2.3	91.2 ± 1.4
AlexNet-RNN [96]	89.7 ± 1.7	-	-
Fus-CNN [101]	84.1 ± 2.7	83.8 ± 2.7	91.3 ± 1.4
Fusion 2D/3D CNNs [105]	89.0 ± 2.1	78.4 ± 2.4	91.8 ± 0.9
STEM-CaRFs [102]	88.8 ± 2.0	80.8 ± 2.1	92.2 ± 1.3
Bu çalışma	89.9 ± 1.6	84.0 ± 1.8	92.5 ± 1.2

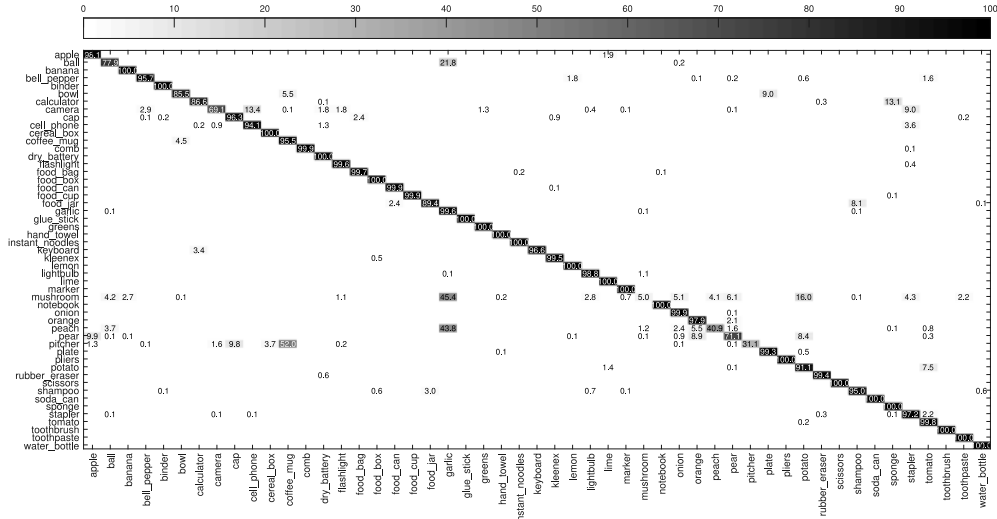
Önerilen yaklaşımın, Washington RGB-D Nesne veri kümesindeki kategoriler için elde ettiği tekil başarılar, Şekil 5.6’da gösterilmektedir. Sonuçlar, önerilen yaklaşımın, çoğu nesne kategorisi için yüksek performans gösterdiğini açıklamaktadır. Genel olarak, daha düşük sonuçlara sahip kategoriler *mushroom* (mantar), *peach* (şeftali) ve *pitcher* (sürahi) sınıflarıdır. Bu durumun ana sebebi, bu kategori sınıflarının veri kümesindeki asgari alt kategori örneklem sayısı olan, yalnızca 3 örneklem içermeleridir. Dolayısıyla, sınıflar arasındaki bu dağılım dengesizliği, öğrenmeyi örneklem sayısı fazla olan kategorilerin lehine desteklemiştir. Ayrıca, veri kümesindeki sınıf-içi çeşitlilik ve sınıflar-arası benzerlik sınıflandırmayı zorlaştıran diğer sebeplerdir. Özellikle, kullanılan Washington RGB-D Nesne veri kümesindeki birçok kategorinin benzerliği, sınıflandırmada karışıklığa yol açmaktadır. Örneğin, veri kümesindeki *ball* (top), *lightbulb* (ampül), *lime* (misket limonu), *pear* (armut), *potato* (patates) ve *tomato* (domates) gibi geometrik olarak benzer birçok kategorinin varlığı, derinlik verilerindeki tanımayı zorlaştırarak başarının düşmesine neden

olmaktadır. Ayrıca, derinlik başarısı düşük olan *camera* (kamera) gibi kategorilerin, parlak yüzeyleri derinlik bilgilerinin bozulmasına ya da eksikliklere sebep olabilir. Bu durum, derinlik başarısını düşürmektedir. RGB veri türünde ise başarı oranı, yukarıda anılan ortak genel sebeplerin yanı sıra, doku bilgisinin zayıf kaldığı *bowl* (kase) ve *plate* (tabak) gibi sınıflarda daha düşüktür. Şekil 5.7, 5.8 ve 5.9’da, sırasıyla derinlik, RGB ve RGB-D verileri için hata matrisleri gösterilmektedir. Hata matrisleri incelendiğinde, yukarıda bahsedilen nedenlerden kaynaklı karışıklıklar olduğu görülmektedir. Sık karıştırılan nesne kategorilerine ait örnekler Şekil 5.10’da görülebilir. Şekilde mavi renkle verilen yukarıdaki kutu, derinlik verileri için sıkça karıştırılan nesne kategorilerine ait RGB görüntüleri örnelemektedir. Görsel olarak uygun olması için derinlik görüntülerine karşılık gelen RGB görüntüleri verilmiştir. Derinlik verilerinin kullanımında hiçbir şekilde RGB verilerinden yararlanılmamıştır. Aşağıdaki turuncu kutuda ise RGB görüntüler için sıkça karıştırılan nesne kategorilerine ilişkin örnekler verilmiştir. Şekil incelendiğinde, yukarıdaki ortak karışıklık nedenlerine (örneklem sayısının eksikliği gibi) ek olarak derinlik verileri için şekilsel olarak karıştırılmaya müsait, benzer örnekleri içeren sınıfların karıştırıldığı, RGB görüntülerde ise doku bilgisinin zayıf kaldığı sınıfların karıştırıldığı görülmektedir.

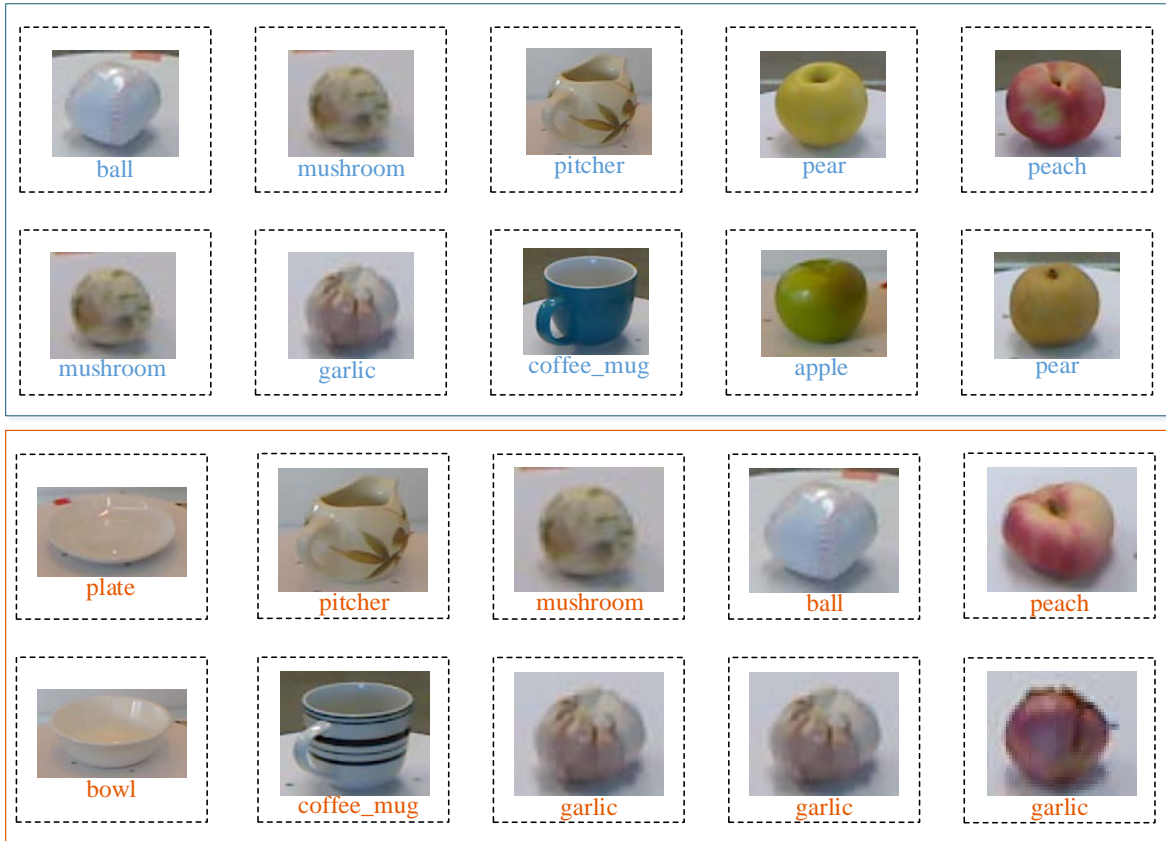


Şekil 5.6. Önerilen yaklaşımın Washington RGB-D Nesne veri kümesindeki kategoriler için tekil sınıflandırma başarıları.





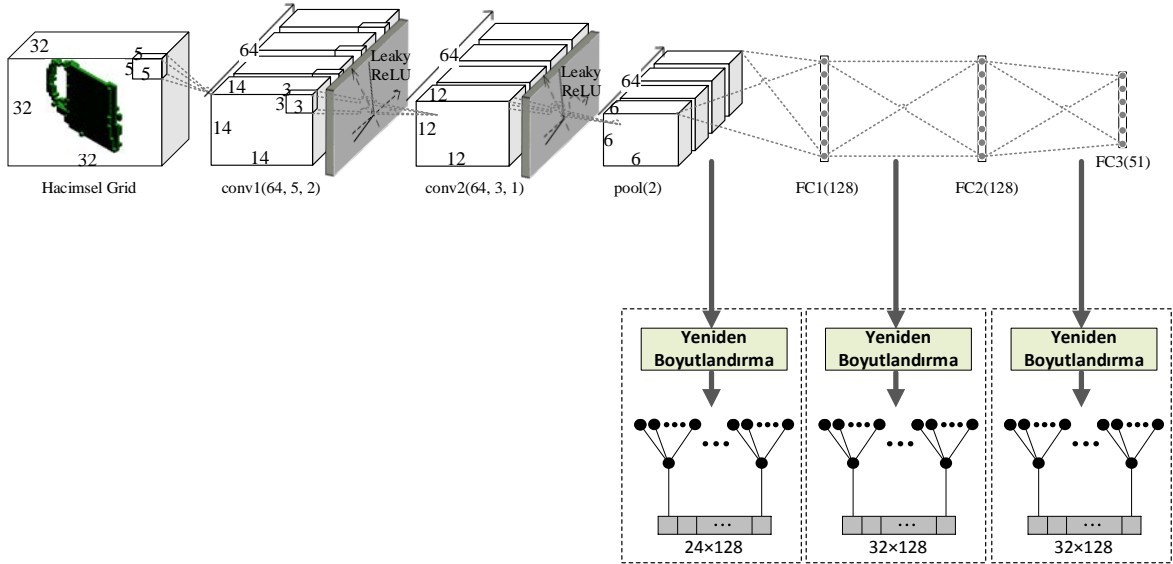
Şekil 5.9. Önerilen yaklaşımın RGB-D verilerinde tanıma hata matrisi (elektronik kopyada zumlayarak, basılı kopyada ise mercekli büyütülerek görüntülenmesi önerilir).



Şekil 5.10. Sıkça karıştırılan nesne kategorilerine ilişkin örnek görüntüler. Yukarıdaki mavi kutu derinlik, turuncu kutu ise RGB verilerinde sıkça karıştırılan kategoriler için örnek RGB görüntülerini göstermektedir. Derinlik verileri kutusunda, RGB görüntüler görsel olarak uygun olması için kullanılmıştır. Derinlik verisinin kullanımında RGB bilgisinden faydalanılmamıştır. Her kutuda üst satırdaki kategoriler, alt satırda karşılık gelen kategori türü ile karıştırılmıştır.

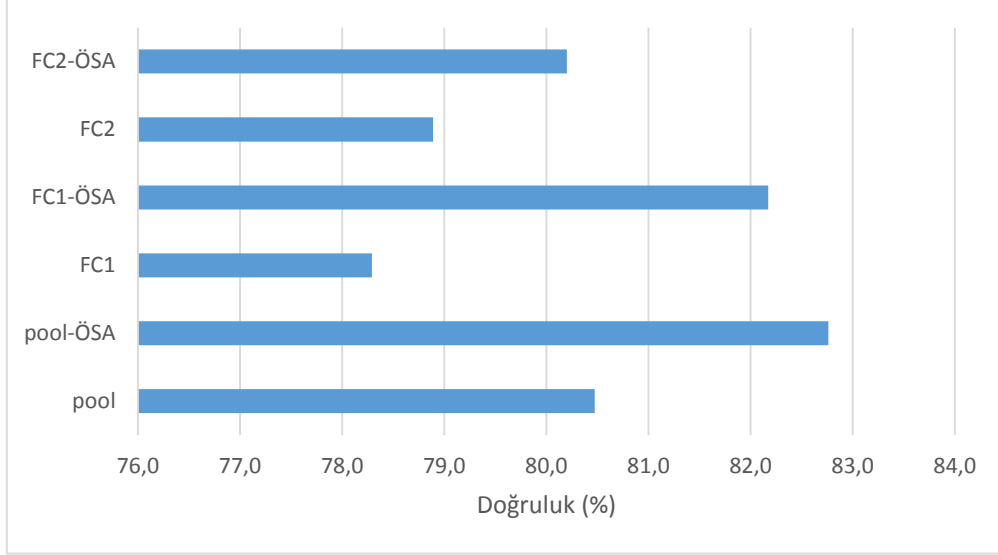
5.4.3. Hacimsel Tanımda ÖSA Modelinin Kullanımı ve Deneysel Sonuçlar

Bu çalışmada ayrıca öz-yinelemeli sinir ağlarının, yukarıda anlatılan RGB ve derinlik verilerine ek olarak, Bölüm 4’te anlatılan hacimsel veri gösterimlerindeki etkisi araştırılmaktadır. Bu amaçla yoğunluk gridleri ve Şekil 4.2’de verilen 3B ESA modeli kullanılmaktadır. 3B ESA modeli için eğitim, Bölüm 4’te anlatıldığı gibi 820 epoch’a kadar yapılmakta ve son katmanda nihai sınıflandırma yapmak yerine, eğitim tamamlandığında aktivasyon ağırlıkları kaydedilmektedir. Erken katmanlardan elde edilen nitelik boyutlarının çok yüksek olmasından dolayı, deneyler havuzlama katmanından itibaren son üç katman çıktıları üzerinde yapılmaktadır. 3B ESA ve ÖSA’nın nasıl uygulandığına dair şematik gösterim Şekil 5.11’de gösterilmektedir. Elde edilen nitelikler, doğrusal SVM ile sınıflandırılmaktadır.



Şekil 5.11. Hacimsel verilerde 3B ESA-ÖSA işbirliğinin kullanımı.

Deneyler için yine Washington RGB-D Nesne veri kümesi kullanıldı. Ancak tüm 10 eğitim/test bölmelerinin kullanımı yerine sadece 2 eğitim/test bölmesi kullanılarak deneyler yapıldı. ÖSA uygulanmadan sadece 3B ESA ile elde edilen başarı oranları ve ÖSA uygulandıktan sonra elde edilen başarı oranları beraber Şekil 5.12’de görülmektedir. Şekilden de görüleceği üzere öz-yinelemeli ağ yapılarının 3B ESA çıktılarına uygulanması ile edilen başarı, her üç katman için de artmıştır. Ancak, nitelik boyutları *FC1* ve *FC2* için ÖSA kullanıldığında doğal olarak artmaktadır. Çünkü bu iki tam-bağlantılı katman çıktıları oldukça küçük olup, 128 iken; ÖSA uygulandıktan sonra $32 \times 128 = 4096$ olmaktadır. Havuzlama katmanı olan *pool* için ise nitelik boyutu $64 \times 6 \times 6 \times 6 = 13,824$ ’ten 3072’e düşürülmektedir.



Şekil 5.12. Hacimsel grid verisi için 3B ESA katmanlarından elde edilen başarı oranlarının, ÖSA uygulanarak elde edilen başarı oranları ile karşılaştırması.

5.5. Sonuç

Bu bölümde, RGB-D nesne kategorilerini tanımak için çoklu ÖSA yapılarının, öneğitimli bir ESA modeli ile birlikte kullanıldığı, başarı performansı yüksek bir derin nitelik çıkartma yaklaşımı sunulmaktadır. ÖSA yapılarının, öneğitimli ESA çıktılarına uygulanması, yüksek boyutlu ESA çıktıları ile baş edilmesini sağlarken, ayrıca daha yüksek doğruluk oranı ile nesne kategorilerini tanımak için farklı katman çıktılarından faydalanması olanağını sunmaktadır.

Derinlik verilerini işlerken, büyük ölçekli RGB veri kümelerinden yararlanabilmek için, derinlik haritaları RGB veri türüne benzer bir yapıda ele alınmaktadır. Bu amaçla derinlik haritalarından yüzey normalleri hesaplanmakta ve her bir boyutu bir renk kanalıymış gibi normalleştirilerek renklendirilmektedir. RGB-D nesne tanıma için yaygın kullanılan bir veri kümesi olan Washington RGB-D Nesne veri kümesinde, çeşitli model parametre seçimleri ve karşılaştırmalı sonuçları, analizleriyle beraber sunulmaktadır. Ayrıca hacimsel grid gösterimleri kullanılarak, kullanılan ÖSA yapılarının 3B ESA modellerinde hacimsel nesne tanımda da sınıflandırma doğruluğu performansını artırdığı gözlenmiştir.

Önerilen yaklaşım, hem nitelik boyutlarının düşürülmesi hem de yüksek sınıflandırma doğruluğu açısından literatürdeki diğer yöntemlere kıyasla başarılı sonuçlar vermektedir. Önerilen yaklaşımın daha da geliştirilmesi için büyük bir potansiyel vardır. Burada etkisi araştırılmayan bir potansiyel faktör, ÖSA ile entegre etmeden önce öneğitimli ESA

modelinin kullanılan veri kümesinde ince ayarlanmasıdır. Özellikle, RGB veri kümesinde eğitilmiş ESA modelinin derinlik verilerinde kullanılması düşünüldüğünde, derinlik verilerindeki ince ayarlama daha iyi sonuçları sağlayabilir. Ayrıca, bu çalışmada önerilen yaklaşımın temelinde VGG-f öneğitilmiş ESA modeli kullanılmaktadır. Son yıllarda önerilmiş daha başarılı sonuçlar veren ResNet [113], DenseNet [114] vb. gibi daha modern öneğitilmiş bir ESA modeli kullanılarak, tanıma doğruluğunu daha da artırmak mümkündür. Bu çalışmada kullanılan ÖSA, rastgele ilklendirilen ağırlıkların kullanıldığı, nitelik temsillerinin eğitim olmaksızın elde edildiği yapılardır. Çoklu ÖSA yapılarını grafik kartlarını kullanarak etkili bir eğitim algoritması ile eğiterek, sonuçları daha da iyileştirmek mümkündür. Son olarak, derinlik verilerinin renklendirilmesinde yüzey normallerinin kullanımı bu çalışmada araştırılmıştır. Başka renklendirme yöntemleri ya da nitelik temsillerinin birleştirilmesi için diğer tekniklerin kullanımı, gelecekte araştırılmaya değer diğer konulardır.

6. SONUÇLAR

Nesne tanıma, bilgisayarlı görü alanında sıkça çalışılan konuların başında gelen temel bir araştırma alanıdır. İnsan-bilgisayar etkileşimi ve robotik alanları başta olmak üzere, çevreyle ya da ortamla etkileşim söz konusu olan problem çözümlerinde, sahnedeki nesnelere tanınması en temel gereksinim olmaktadır. Derinlik görüntülerini RGB görüntülerle beraber eşzamanlı sağlayan Kinect gibi RGB-D kamera teknolojisinin gelişmesiyle beraber, RGB-D verilerinin kullanımı, robotik görme alanı başta olmak üzere birçok alanda yaygınlaşmıştır. RGB-D nesne tanıma temel bir problem olarak sıkça çalışılan bu konuların başında gelmektedir. Öte yandan, makine öğrenmesinin bir alt dalı olarak gelişen ve bugün tek başına bir alan olan derin öğrenme tekniklerindeki gelişmeler, çeşitli tanıma problemlerinde büyük bir performans artışı sağlamıştır. Bu tez kapsamında, derin öğrenme tekniklerini kullanarak RGB-D verilerinde nesne kategorilerini tanımak üzere, üç farklı çalışma yapılmıştır. Bu çalışmalarda, evrişimsel sinir ağları ve öz-yinelemeli sinir ağları olmak üzere iki derin öğrenme tekniği kullanılmıştır.

Tezin ilk kesiminde, derin niteliklerin eğitim gerektirmeksizin elde edildiği, sık bir mimari üzerinde deneysel bir analiz çalışması yapılmıştır. Kullanılan mimari, tek bir ESA katmanı ve ESA katmanından elde edilen çıktılar giriş olarak kabul eden çoklu ÖSA yapılarından oluşmaktadır. Temel alınan mimari, geri-yayılım algoritması kullanmadan derin nitelikleri elde eden RGB-D nesne tanıma çalışmalarında, sıkça kullanılan bir modeldir. Ancak bu çalışmalarda, RGB ve derinlik verilerinin kullanımında aynı model ayarlamaları temel alınmaktadır. Öte yandan, RGB ve derinlik verilerinin karakteristikleri farklı olduğundan dolayı, bu durum, Bölüm 3'te anlatılan çalışmada, deneysel bir analiz ile sorgulanmaktadır. Deneysel, beklendiği gibi farklı model ayarlamalarının RGB ve derinlik verilerinde ayrı ayrı daha iyi sonuçlar verdiğini doğrulamaktadır. RGB verilerinde maksimum havuzlama ve mutlak değerleri alan doğrultucu birim fonksiyonu en iyi sonuçları üretirken, derinlik için ortalama havuzlama ve *ReLU/leaky ReLU* doğrultucu birim fonksiyonları daha iyi sonuçlar sağlamıştır. Rastgele yamalardan öğrenilen evrişim filtreleri ile SIFT/SURF ilgi noktaları etrafından öğrenilen filtreler arasında küçük bir fark gözlenerek, rastgele çıkartılan görüntü yamalarından etkili evrişim ağırlıkları öğrenildiği sonucuna varılmıştır. Sonuç olarak, bu çalışma sonucunda uygun model ayarlamaları ile RGB-D nesne tanıma doğruluğunun ~2% civarında artırılabilirdiği gösterilmiştir. Son yıllarda, büyük ölçekli veri kümelerinde, gradyan tabanlı derin öğrenme çalışmaları yaygınlaştıkça da, gözetimsiz

öğrenilen filtre ağırlıkları ile evrişimsel nesne tanıma, özellikle küçük ölçekli veri kümelerinde yararlı olabilir.

Derinlik verileri; (i) derinlik haritaları olarak tek kanallı görüntü yapıları, (ii) üç boyutlu uzayda ifade edilen nokta bulutu, (iii) yüzey ve kenar bilgilerinin tutulduğu mesh modelleri ve (iv) hacimsel grid temsilleri gibi farklı veri temsillerinde ele alınabilir. Tez çalışmaları kapsamında sonraki kesimde (Bölüm 4), derinlik verisinde saklı olan, nesnelere geometrik yapısını daha iyi ortaya çıkartmak için derinlik verileri, hacimsel temsiller ile ele alınmıştır. Bu amaçla, derinlik görüntülerini kullanan ikili grid ve yoğunluk gridi olmak üzere iki farklı hacimsel temsil önerilmiştir. Daha sonra tanımlanan hacimsel temsilleri girdi olarak ele alan, 3B ESA ile öğrenme gerçekleştirilmiştir. Hacimsel gridler, tek-dönümlü ve çoklu-dönümlü yaklaşımları ile ele alınmıştır. Deneysel sonuçlar, yoğunluk gridinin ikili gride göre daha başarılı bir temsil olduğunu doğrularken, çoklu-dönümlü yaklaşım ile elde edilen başarının tek-dönümlü yaklaşım başarısını geçtiğini göstermiştir. Ancak çoklu-dönümlü yaklaşımda, birden fazla nesne temsili, girdi olarak ele alındığı için tek-dönümlü yaklaşımı ile doğrudan bir karşılaştırma yapmak adil olmayabilir. Bu amaçla, tek bir giriş temsili için dönme matrisleri ile z-ekseni boyunca döndürerek çoğaltan, çoklayan-dönümlü yaklaşımı önerilmiştir. Bu yaklaşım ile küçük veri kümesi olan 2D3D Nesne veri kümesinde başarı derecesi tek-dönümlü yaklaşıma göre önemli derecede artırılmıştır. Ancak, daha büyük veri kümesi olan Washington RGB-D veri kümesinde başarı, küçük veri kümesinde olduğu gibi artmamıştır. Bu bölüm kapsamında son olarak renk bilgisinin hacimsel gösterimde kodlandığı yaklaşımlar önerilmiştir. Bu amaçla, gri-tonlamalı grid, 8-bit renk gridi, hiperküp gridi ve RGB çoklu-dönümlü yaklaşımı olmak üzere dört farklı yaklaşım önerilmiştir. Deneysel sonuçlar, gri-tonlamalı grid ile RGB çoklu-dönümlü yaklaşımlarının sadece derinlik bilgisinin kullanıldığı duruma göre başarıyı artırdığını, ancak 8-bit renk gridi ile hiperküp gridi yaklaşımlarının başarıyı düşürdüğünü göstermiştir. Fazladan renk bilgilerinden faydalanmasına rağmen başarı performansının istenen düzeyde olmaması, uygun bir 3B ESA mimari yapısının kullanılmamasından kaynaklı olabilir. Bu amaçla, ileriki çalışmalarda deneysel olarak farklı mimariler kullanılarak, uygun bir model bulunması hedeflenmektedir. Öte yandan, hacimsel temsillerin kullanımının dezavantajları da vardır. Bunların başlıcası, yüksek boyutlarla (*curse of dimensionality*) başetmektir. Örneğin, $30 \times 30 \times 30$ 'luk bir hacimsel tensörün 2B'deki eşdeğeri yaklaşık olarak 165×165 olmaktadır ($30 \ll 165$). Ayrıca bu yüksek boyutlar içerisinde, nesne küçük bir alanı kaplarken, büyük boşluklar olabilmektedir. Bu

dezavantajları aşmak için, [115] çalışmasında hacimsel temsillerden 2B temsilleri elde edilerek daha yüksek bir başarı sağlanmıştır. Ancak tez kapsamında kullanılan görüntü açısına bağlı eksik modeller yerine, bu çalışmada [115], ModelNet CAD modelleri kullanılmaktadır. Bu yüzden, aynı anlayışın eksik veri modellerinde başarı elde etmesi şüphelidir. Öte yandan, [116] çalışmasında, eşyönlü (*isotropic*) filtre kullanımına alternatif olarak, eşyönsüz (*anisotropic*) yaklaşımı önerilerek 3B modellerinden 2B modellerine geçiş yapılmaktadır. Benzer bir anlayış ile bu tez kapsamında sunulan çalışmalar geliştirilebilir.

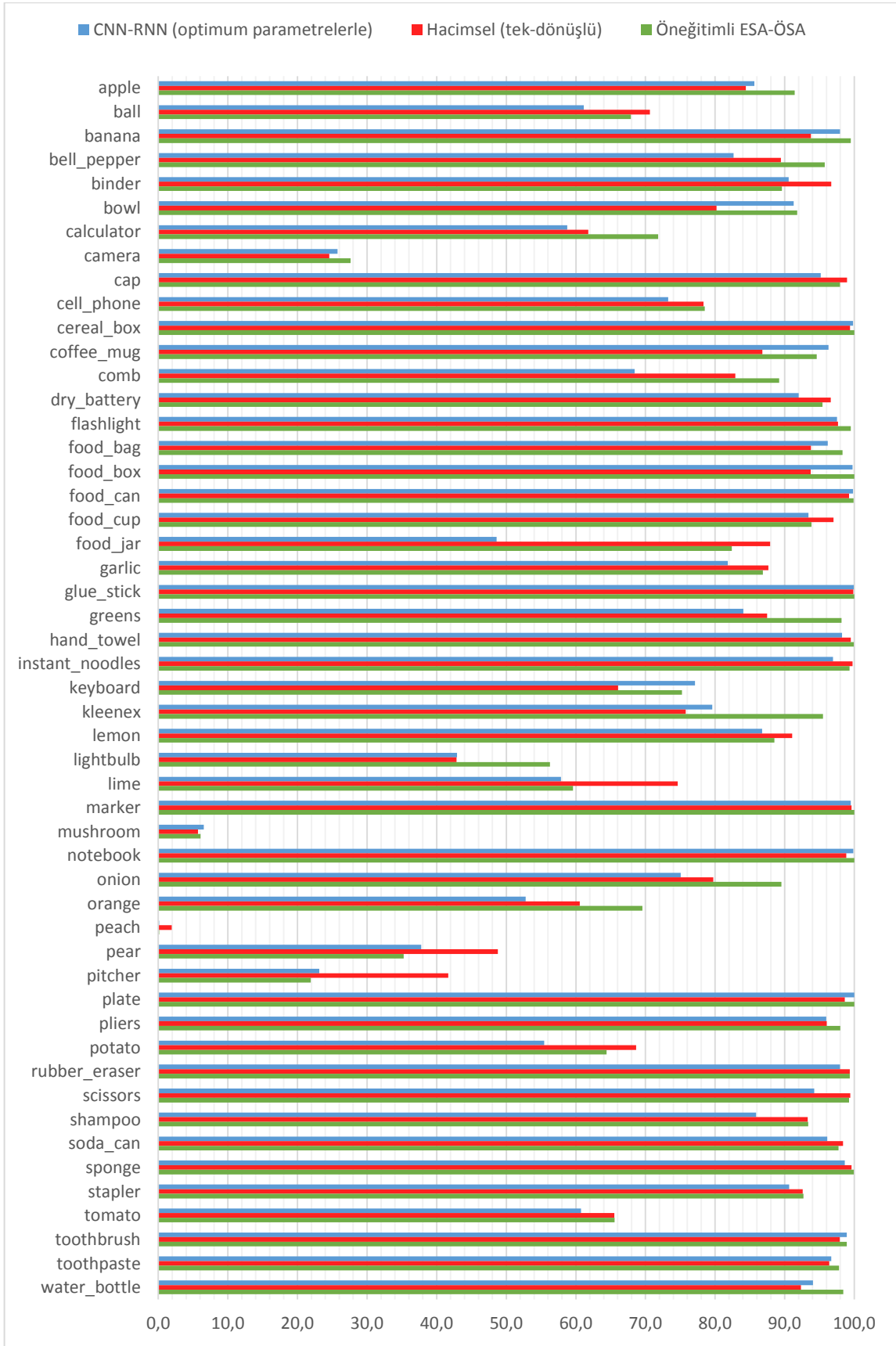
Tezin sonraki kesimi Bölüm 5'te, transfer öğrenme ile RGB-D nesne tanıma çalışması gerçekleştirilmiştir. Bu amaçla öneğitimli bir ESA modeli, önerilen yaklaşımın temelinde kullanılırken, ESA çıktılarını giriş olarak ele alan çoklu ÖSA modelleri ile üst düzey nitelik temsilleri elde edilmektedir. Derinlik verileri için geniş ölçekli RGB veri kümelerinden yararlanabilmek için, bu çalışmada, derinlik haritalarından hesaplanan yüzey normalleri, RGB renk kanallarına benzer şekilde ele alınarak 0 – 255 değerlerine normalleştirilerek renklendirilmektedir. Öneğitimli ESA'nın farklı katmanlardan elde edilen aktivasyon çıktıları, ÖSA yapılarına uygun olarak yeniden boyutlandırılarak, giriş olarak ÖSA'ya verilmektedir. Deneysel olarak orta katman çıktılarının temsil kabiliyeti daha yüksek nitelikleri sundukları gözlenerek, orta düzey nitelik temsillerinin birleştirilmeleri ile nesne görüntülerinin tamamını güçlü bir şekilde temsil eden vektörler elde edilmiştir. Nihai vektörlerin doğrusal SVM sınıflandırıcısı ile sınıflandırılmalarıyla, literatürdeki ilgili diğer çalışmalar aşılıp, hem tanıma doğruluğu hem de nitelik boyutlarının küçüklüğü açısından etkili bir yaklaşım sunulmuştur. Önerilen bu yaklaşımın nitelik çıkartma aşaması, eğitim olmaksızın hızlı bir şekilde çalışabilmektedir. Bu yaklaşım, RGB-D nesne tanıma problemi yanı sıra, tespit etme, semantik bölütleme ve eylem tanıma gibi derin niteliklerin kullanılabilmesi için çeşitli görme problemlerine kolaylıkla uygulanabilir. Özellikle niteliklerin küçük boyutlarda eğitim gerektirmeksizin hızlı bir şekilde elde edilmesi, gerçek zamanlı robotik uygulamalarında yararlı olabilir. Bu çalışmada, öneğitimli ESA modeli olarak VGG-f olarak adlandırılan basit bir sıradüzensel ESA modeli kullanılmaktadır. Bu model, son yıllarda önerilen ResNet [113], DenseNet [114] gibi daha modern ESA modelleri ile değiştirilerek daha başarılı sonuçlar elde etmek mümkündür. Ayrıca kullanılan veri kümesinde öneğitimli ESA modeli ince-ayar yapılarak ve etkili bir öğrenme algoritması ile ÖSA yapıları eğitilerek başarıyı daha da artırmak mümkündür. Bu durumda, elbette eğitim kaynaklı fazladan zaman maliyeti ortaya çıkar.

Bunların yanı sıra, farklı düzey çıktılarının birleştirilmeleri aşamasında, daha etkili başka nitelik birleştirme teknikleri araştırılabilir.

6.1. Tartışmalar ve Gelecek Çalışmalar

Bu tez kapsamında, nesne kategorilerini RGB-D verileri ile tanıma problemine odaklanılmıştır. Geriyayılım algoritması kullanılmaksızın derin niteliklerin elde edilmesi ile tanıma, 3B ESA ile derinlik verilerinde hacimsel tanıma ve transfer öğrenme ile tanıma olmak üzere üç farklı yaklaşım ile problem ele alınmıştır. Derin öğrenme tekniklerinden ESA ve çoklu ÖSA ağaç yapıları, önerilen yöntemlerde kullanılmıştır. Tezin ilk kesiminde konu edilen yaklaşıma göre, derinlik verileri RGB renk kanallarına ek bir kanalmış gibi ele alınmaktadır. Ayrıca gradyan tabanlı olmayan bu yaklaşımda geriyayılım algoritması kullanılmadan nitelikler elde edilmektedir. Dolayısıyla, hem derinlik verisinin doğasına daha uygun bir yaklaşım ile verileri ele almak, hem de gradyan tabanlı öğrenme anlayışı ile nesne tanıma başarısını artırmak amaçlarıyla sonraki kesimde hacimsel nesne tanıma çalışmaları gerçekleştirilmiştir. Amaçlandığı gibi başarı derecesi, derinlik verileri için önemli oranda artarak 79.7%'den 82.4%'e çıkmıştır. Ancak önerilen bu hacimsel tanıma yaklaşımında RGB-D veri kümelerinde derinlik verileri kullanılarak eğitim sıfırdan gerçekleştirilmektedir. RGB veri türü için, ImageNet gibi milyonlarca veri içeren geniş ölçekli veri kümeleri mevcuttur. Ancak derinlik verisi için bu denli geniş çaplı bir veri kümesi henüz tanımlanmamıştır. Derin öğrenme ağları, veriye aç modellerdir. Bu yüzden öğrenmenin geniş ölçekli veri kümeleri üzerinde yapılması, sistemin başarımını doğrudan etkiler. Dolayısıyla, tezin sonraki kesiminde geniş ölçekli ImageNet veri kümesinde eğitilmiş hazır bir ESA modelini çoklu ÖSA yapıları ile birlikte kullanarak, etkin bir transfer öğrenme tabanlı yaklaşım önerilmiştir. Bu yeni yaklaşım, ImageNet'in veri büyüklüğünden yararlanarak, derinlik verisi için başarı derecesini bir önceki kesimde yapılan çalışmaya göre önemli oranda artırarak 82.4%'ten 84.0%'a çıkartmıştır. Her üç yaklaşımın, Washington RGB-D Nesne veri kümesinde derinlik verilerini kullanarak elde ettikleri, tekil kategori başarıları karşılaştırmalı olarak Şekil 6.1'de gösterilmektedir. Verilen grafik incelendiğinde, üçüncü bölümde anlatılan çalışma, 51 kategorinin sadece 4'ü için diğerlerine göre daha iyi sonuç üretirken, dördüncü bölümde sunulan hacimsel yaklaşımla nesne tanıma yaklaşımı, 16 kategori için en iyi sonucu üretmektedir. Tezin son kesiminde ele alınan transfer öğrenme ile nesne tanıma yaklaşımı ise, 51 kategorinin 31'i için diğer çalışmalara göre daha iyi sonucu üreterek en başarılı yaklaşım olmuştur.

Gelecek çalışmalar için, semantik bölütleme, tespit etme (*detection*) gibi diğer konuların ele alınması planlanmaktadır. Ayrıca temelinde nesnelere tanımanın olduğu, semantik haritalama, bir sahnedeki olayı tanıma gibi daha bütüncül bakış açısı ile ele alınabilecek robotik görüş sistemlerine katkıda bulunması hedeflenmektedir. RGB-D nesne tanıma problemi özelinde, gözetimsiz öğrenme veya küçük veri kümelerinde etkin öğrenme yöntemleri araştırılması planlanmaktadır. Tez kapsamında, sıkça karşılaşılan veri kümesindeki dağılım dengesizliğinden kaynaklanan yanlı öğrenme (*biased learning*) problemlerine ilişkin, çekişmeli üretici ağlar (*generative adversarial networks*) [117] ile alt kategori örnekleme az olan nesnelere için yeni veriler üretilerek daha dengeli bir öğrenme sağlanabilir. Son olarak, ESA'ların nitelikleri hiyerarşik bir şekilde öğrenirlerken, öteleme (*translation*) ve dönme (*rotation*) gibi poz bilgilerini dikkate almayan eksikliklerini tamamlamak üzere geliştirilen, kapsül ağların [118], [119] kullanımı, RGB-D nesne tanıma için araştırılabilir.



Şekil 6.1. Önerilen yaklaşımlar ile Washington RGB-D Nesne veri kümesinde derinlik verilerini kullanarak, tekil kategoriler için elde edilen sınıflandırma başarıları (%).

KAYNAKLAR

- [1] Andreopoulos, A., Tsotsos, J. K., 50 years of object recognition: Directions forward, *Computer vision and image understanding* vol. 117, no. 8, 827--891 Elsevier, **2013**.
- [2] Marr, D., Vision: The philosophy and the approach, **1982**.
- [3] Pinto, N., Cox, D. D., DiCarlo, J. J., Why is Real-World Visual Object Recognition Hard?, *PLoS Computational Biology*, 4, e27, **2008**.
- [4] Torralba, A., Efros, A. A., Unbiased look at dataset bias, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1521--1528, **2011**.
- [5] Hubel, D. H., Wiesel, T. N., Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *The Journal of physiology*, 160, 106--154, **1962**.
- [6] Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A.J. and Wiskott, L., Deep Hierarchies in the Primate Visual Cortex : What Can We Learn For Computer Vision ?, *IEEE transactions on pattern analysis and machine intelligence*, 35, 1847--1871, **2013**.
- [7] Marr, D., Vision: A computational investigation into the human representation and processing of visual information. MIT Press, Cambridge, Massachusetts, **1982**.
- [8] LeCun, Y., Bengio, Y., Hinton, G., Deep learning, *nature*, 521, 436, **2015**.
- [9] Sarbolandi, H., Lefloch, D., Kolb, A., Kinect range sensing: Structured-light versus Time-of-Flight Kinect, *Computer vision and image understanding*, 139, 1--20, **2015**.
- [10] Han, J., Shao, L., Xu, D., Shotton, J., Enhanced computer vision with microsoft kinect sensor: A review, *IEEE transactions on cybernetics*, 43, 1318--1334, **2013**.
- [11] Alexe, B., Deselaers, T., Ferrari, V., What is an object?, *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 73--80, **2010**.
- [12] Caglayan, A., Can, A. B., An Empirical Analysis of Deep Feature Learning for RGB-D Object Recognition, *International Conference Image Analysis and Recognition*, 312--320, **2017**.
- [13] Caglayan, A., Can, A. B., 3D convolutional object recognition using volumetric representations of depth data, *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*, 125--128, **2017**.
- [14] Caglayan, A., Can, A. B., Volumetric Object Recognition Using 3-D CNNs on Depth Data, *IEEE Access*, 6, 20058--20066, **2018**.
- [15] Caglayan, A., Can, A. B., Exploiting Multi-Layer Features Using a CNN-RNN Approach for RGB-D Object Recognition, *Computer Vision - ECCV 2018 Workshops*, Cham, **2018**.
- [16] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M. S., Deep learning for visual understanding: A review, *Neurocomputing*, 187, 27--48, **2016**.
- [17] Hubel, D. H., Wiesel, T. N., Receptive fields and functional architecture of monkey striate cortex, *The Journal of physiology*, 195, 215--243, **1968**.
- [18] Fukushima, K., Miyake, S., Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, *Competition and cooperation in*

- neural nets*, Springer, 267–285, **1982**.
- [19] Rumelhart, D. E., Hinton, G. E., Williams, R. J., Learning representations by back-propagating errors, *nature*, 323, 533, **1986**.
- [20] LeCun, Y. *et al.*, Handwritten digit recognition with a back-propagation network, *Advances in neural information processing systems*, 396–404, **1990**.
- [21] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86, 2278–2324, **1998**.
- [22] Krizhevsky, A., Sutskever, I., Hinton, G. E., Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 1097–1105, **2012**.
- [23] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., Imagenet: A large-scale hierarchical image database, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255, **2009**.
- [24] Geman, S., Bienenstock, E., Doursat, R., Neural networks and the bias/variance dilemma, *Neural computation*, 4, 1–58, **1992**.
- [25] Goodfellow, I., Bengio, Y., Courville, A., Deep Learning, vol. 1. *Cambridge: MIT press*, **2016**.
- [26] Wu, Z. *et al.*, 3d shapenets: A deep representation for volumetric shapes, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920, **2015**.
- [27] Lin, M., Chen, Q., Yan, S., Network in network, *CoRR*, abs/1312.4, **2013**.
- [28] Sutskever, I., Martens, J., Dahl, G., Hinton, G., On the importance of initialization and momentum in deep learning, *International conference on machine learning*, 1139–1147, **2013**.
- [29] Glorot, X., Bengio, Y., Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256, **2010**.
- [30] He, K., Zhang, X., Ren, S., Sun, J., Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, 1026–1034, **2015**.
- [31] Mishkin, D., Matas, J., All you need is a good init, *CoRR*, abs/1511.0, **2015**.
- [32] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, 15, 1929–1958, **2014**.
- [33] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R., Regularization of neural networks using dropconnect, *International Conference on Machine Learning*, 1058–1066, **2013**.
- [34] Zou, H., Hastie, T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320, **2005**.
- [35] Ren, S., He, K., Girshick, R., Sun, J., Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1137–1149, **2017**.

- [36] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., You only look once: Unified, real-time object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788, **2016**.
- [37] Li, H., Li, Y., Porikli, F., Deeptack: Learning discriminative feature representations online for robust visual tracking, *IEEE Transactions on Image Processing*, 25, 1834–1848, **2016**.
- [38] Hong, S., You, T., Kwak, S., Han, B., Online tracking by learning discriminative saliency map with convolutional neural network, *International Conference on Machine Learning*, 597–606, **2015**.
- [39] Huang, W., Qiao, Y., Tang, X., Robust scene text detection with convolution neural network induced msr trees, *European Conference on Computer Vision*, 497–511, **2014**.
- [40] He, P., Huang, W., Qiao, Y., Loy, C. C., Tang, X., Reading Scene Text in Deep Convolutional Sequences., *AAAI*, 3501–3508, 16, **2016**.
- [41] Zhang, Y., Cheng, L., Wu, J., Cai, J., Do, M. N., Lu, J., Action recognition in still images with minimum annotation efforts, *IEEE Transactions on Image Processing*, 25, 5479–5490, **2016**.
- [42] Ji, S., Xu, W., Yang, M., Yu, K., 3D convolutional neural networks for human action recognition, *IEEE transactions on pattern analysis and machine intelligence*, 35, 221–231, **2013**.
- [43] Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., Yu, D., Convolutional neural networks for speech recognition, *IEEE/ACM Transactions on audio, speech, and language processing*, 22, 1533–1545, **2014**.
- [44] Yu, D. *et al.*, Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention., *Interspeech*, 17–21, **2016**.
- [45] Kim, Y., Convolutional Neural Networks for Sentence Classification, *CoRR*, abs/1408.5, **2014**.
- [46] Dauphin, Y. N., Fan, A., Auli, M., Grangier, D., Language Modeling with Gated Convolutional Networks, *CoRR*, abs/1612.0, **2016**.
- [47] Gu, J. *et al.*, Recent advances in convolutional neural networks, *Pattern Recognition*, 77, 354–377, **2018**.
- [48] Frasconi, P., Gori, M., Sperduti, A., A general framework for adaptive processing of data structures, *IEEE transactions on Neural Networks*, 9, 768–786, **1998**.
- [49] Socher, R., Manning, C. D., Ng, A. Y., Learning continuous phrase representations and syntactic parsing with recursive neural networks, *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, **2010**.
- [50] Pollack, J. B., Recursive distributed representations, *Artificial Intelligence*, 46, 77–105, **1990**.
- [51] Hinton, G. E., Mapping part-whole hierarchies into connectionist networks, *Artificial Intelligence*, 46, 47–75, **1990**.
- [52] Socher, R., Lin, C. C.-Y., Ng, A. Y., Manning, C. D., Parsing natural scenes and natural language with recursive neural networks, *Proceedings of the 28th international conference on machine learning (ICML-11)*, 129–136, **2011**.

- [53] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., Manning, C. D., Semi-supervised recursive autoencoders for predicting sentiment distributions, *Proceedings of the conference on empirical methods in natural language processing*, 151–161, **2011**.
- [54] Socher, R., Huang, E. H., Pennin, J., Manning, C. D., Ng, A. Y., Dynamic pooling and unfolding recursive autoencoders for paraphrase detection, *Advances in neural information processing systems*, 801–809, **2011**.
- [55] Socher, R. *et al.*, Recursive deep models for semantic compositionality over a sentiment treebank, *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642, **2013**.
- [56] Socher, R., Huval, B., Bath, B., Manning, C. D., Ng, A. Y., Convolutional-recursive deep learning for 3d object classification, *Advances in neural information processing systems*, 656–664, **2012**.
- [57] Jarrett, K., Kavukcuoglu, K., LeCun, Y., others, What is the best multi-stage architecture for object recognition?, *Computer Vision, 2009 IEEE 12th International Conference on*, 2146–2153, **2009**.
- [58] Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., Ng, A. Y., On Random Weights and Unsupervised Feature Learning., *ICML*, 1089–1096, **2011**.
- [59] Coates, A., Ng, A., Lee, H., An analysis of single-layer networks in unsupervised feature learning, *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223, **2011**.
- [60] Blum, M., Springenberg, J. T., Wülfing, J., Riedmiller, M. A., A learned feature descriptor for object recognition in rgb-d data., *ICRA*, 1298–1303, **2012**.
- [61] Bai, J., Wu, Y., Zhang, J., Chen, F., Subset based deep learning for RGB-D object recognition, *Neurocomputing*, 165, 280–292, **2015**.
- [62] Bo, L., Ren, X., Fox, D., Unsupervised feature learning for RGB-D based object recognition, *Experimental Robotics*, 387–402, **2013**.
- [63] Cheng, Y., Zhao, X., Huang, K., Tan, T., Semi-supervised learning for rgb-d object recognition, *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2377–2382, **2014**.
- [64] Cheng, Y., Zhao, X., Huang, K., Tan, T., Semi-supervised learning and feature evaluation for RGB-D object recognition, *Computer Vision and Image Understanding*, 139, 149–160, **2015**.
- [65] Guo, Q., Wang, F., Lei, J., Tu, D., Li, G., Convolutional feature learning and hybrid CNN-HMM for scene number recognition, *Neurocomputing*, 184, 78–90, **2016**.
- [66] Jhuo, I.-H., Gao, S., Zhuang, L., Lee, D. T., Ma, Y., Unsupervised feature learning for RGB-D image classification, *Asian Conference on Computer Vision*, 276–289, **2014**.
- [67] Lai, K., Bo, L., Ren, X., Fox, D., A large-scale hierarchical multi-view rgb-d object dataset, *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 1817–1824, **2011**.
- [68] Lowe, D. G., Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, 60, 91–110, **2004**.

- [69] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., Speeded-up robust features (SURF), *Computer vision and image understanding*, 110, 346–359, **2008**.
- [70] Maas, A. L., Hannun, A. Y., Ng, A. Y., Rectifier nonlinearities improve neural network acoustic models, *Proc. ICML*, 30, **2013**.
- [71] Miller, G., *WordNet: An electronic lexical database*, MIT press, **1998**.
- [72] Bo, L., Ren, X., Fox, D., Depth kernel descriptors for object recognition, *IEEE International Conference on Intelligent Robots and Systems*, 821–826, **2011**.
- [73] Nevatia, R., Binford, T. O., Description and recognition of curved objects, *Artificial Intelligence*, 8, 77–98, **1977**.
- [74] Marr, D., Nishihara, H. K., Representation and recognition of the spatial organization of three-dimensional shapes, *Proc. R. Soc. Lond. B*, 200, 269–294, **1978**.
- [75] Brooks, R. A., Symbolic reasoning among 3-D models and 2-D images., *Artif. Intell.*, 17, 285–348, **1981**.
- [76] Newcombe, R. A. *et al.*, KinectFusion: Real-time dense surface mapping and tracking, *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, 127–136, **2011**.
- [77] Maturana, D., Scherer, S., Voxnet: A 3d convolutional neural network for real-time object recognition, *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 922–928, **2015**.
- [78] Garcia-Garcia, A., Gomez-Donoso, F., Garcia-Rodriguez, J., Orts-Escolano, S., Cazorla, M., Azorin-Lopez, J., Pointnet: A 3d convolutional neural network for real-time object class recognition, *Neural Networks (IJCNN), 2016 International Joint Conference on*, 1578–1584, **2016**.
- [79] Gomez-Donoso, F., Garcia-Garcia, A., Garcia-Rodriguez, J., Orts-Escolano, S., Cazorla, M., Lonchanet: A sliced-based cnn architecture for real-time 3d object recognition, *2017 International Joint Conference on Neural Networks (IJCNN)*, 412–418, **2017**.
- [80] Browatzki, B., Fischer, J., Graf, B., Bühlhoff, H. H., Wallraven, C., Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset, *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 1189–1195, **2011**.
- [81] Bo, L., Lai, K., Ren, X., Fox, D., Object recognition with hierarchical kernel descriptors, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1729–1736, **2011**.
- [82] Tang, S. *et al.*, Histogram of oriented normal vectors for object recognition with a depth sensor, *Asian conference on computer vision*, 525–538, **2012**.
- [83] Zaki, H. F. M., Shafait, F., Mian, A., Convolutional hypercube pyramid for accurate RGB-D object category and instance recognition, *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 1685–1692, **2016**.
- [84] Dalal, N., Triggs, B., Histograms of oriented gradients for human detection, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 886–893, 1, **2005**.

- [85] Johnson, A. E., Hebert, M., Using spin images for efficient object recognition in cluttered 3D scenes, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 433–449, **1999**.
- [86] Lazebnik, S., Schmid, C., Ponce, J., Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2169–2178, **2006**.
- [87] Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., Betz, M., Towards 3D point cloud based object maps for household environments, *Robotics and Autonomous Systems*, 56, 927–941, **2008**.
- [88] Dieleman, S. *et al.*, Lasagne: first release, *Zenodo: Geneva, Switzerland*, 3, **2015**.
- [89] Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., CNN features off-the-shelf: an astounding baseline for recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813, **2014**.
- [90] Schwarz, M., Schulz, H., Behnke, S., RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features, *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, 1329–1335, **2015**.
- [91] Girshick, R., Donahue, J., Darrell, T., Malik, J., Rich feature hierarchies for accurate object detection and semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587, **2014**.
- [92] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, *CoRR*, abs/1312.6, **2013**.
- [93] Farabet, C., Couprie, C., Najman, L., LeCun, Y., Learning hierarchical features for scene labeling, *IEEE transactions on pattern analysis and machine intelligence*, 35, 1915–1929, **2013**.
- [94] Yosinski, J., Clune, J., Bengio, Y., Lipson, H., How transferable are features in deep neural networks?, *Advances in neural information processing systems*, 3320–3328, **2014**.
- [95] Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., Hypercolumns for object segmentation and fine-grained localization, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 447–456, **2015**.
- [96] Bui, H. M., Lech, M., Cheng, E., Neville, K., Burnett, I. S., Object recognition using deep convolutional features transformed by a recursive network structure, *IEEE Access*, 4, 10059–10066, **2016**.
- [97] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., Return of the Devil in the Details: Delving Deep into Convolutional Nets, *CoRR*, abs/1405.3, **2014**.
- [98] Azizpour, H., Sharif Razavian, A., Sullivan, J., Maki, A., Carlsson, S., From generic to specific deep representations for visual recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 36–45, **2015**.
- [99] Oquab, M., Bottou, L., Laptev, I., Sivic, J., Learning and transferring mid-level image representations using convolutional neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724, **2014**.
- [100] Gupta, S., Girshick, R., Arbeláez, P., Malik, J., Learning rich features from RGB-D images for object detection and segmentation, *European Conference on Computer*

- Vision*, 345–360, **2014**.
- [101] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., Burgard, W., Multimodal deep learning for robust rgb-d object recognition, *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 681–687, **2015**.
- [102] Asif, U., Bennamoun, M., Sohel, F. A., Rgb-d object recognition and grasp detection using hierarchical cascaded forests, *IEEE Transactions on Robotics*, 33, 547–564, **2017**.
- [103] Simonyan, K., Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, abs/1409.1, **2014**.
- [104] Cheng, Y., Cai, R., Zhao, X., Huang, K., Convolutional fisher kernels for rgb-d object recognition, *3D Vision (3DV), 2015 International Conference on*, 135–143, **2015**.
- [105] Zia, S., Yüksel, B., Yüret, D., Yemez, Y., RGB-D object recognition using deep convolutional neural networks, *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 887–894, **2017**.
- [106] Sharma, A., Tuzel, O., Liu, M.-Y., Recursive context propagation network for semantic scene labeling, *Advances in Neural Information Processing Systems*, 2447–2455, **2014**.
- [107] Kim, J., Kwon Lee, J., Mu Lee, K., Deeply-recursive convolutional network for image super-resolution, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1645, **2016**.
- [108] Vedaldi, A., Lenc, K., Matconvnet: Convolutional neural networks for matlab, *Proceedings of the 23rd ACM international conference on Multimedia*, 689–692, **2015**.
- [109] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., LIBLINEAR: A library for large linear classification, *Journal of machine learning research*, 9, 1871–1874, **2008**.
- [110] Zeiler, M. D., Fergus, R., Visualizing and understanding convolutional networks, *European conference on computer vision*, 818–833, **2014**.
- [111] Wold, S., Esbensen, K., Geladi, P., Principal component analysis, *Chemometrics and intelligent laboratory systems*, 2, 37–52, **1987**.
- [112] Liu, H., Li, F., Xu, X., Sun, F., Multi-modal local receptive field extreme learning machine for object recognition, *Neurocomputing*, 277, 4–11, **2018**.
- [113] He, K., Zhang, X., Ren, S., Sun, J., Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, **2016**.
- [114] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q., Densely Connected Convolutional Networks, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2017**.
- [115] Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., Multi-View Convolutional Neural Networks for 3D Shape Recognition, *Proceedings of the IEEE International Conference on Computer Vision*, 945–953, **2015**.
- [116] Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L. J., Volumetric and

- Multi-View CNNs for Object Classification on 3D Data, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5648–5656, **2016**.
- [117] Goodfellow, I. *et al.*, Generative Adversarial Nets, *Advances in Neural Information Processing Systems*, 2672–2680, **2014**.
- [118] Sabour, S., Frosst, N., Hinton, G. E., Dynamic routing between capsules, *Advances in Neural Information Processing Systems*, 3856–3866, **2017**.
- [119] Hinton, G. E., Sabour, S., Frosst, N., Matrix capsules with EM routing, *Proc. Int. Conf. Learn. Representations (ICLR)*, **2018**.

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Ali Çağlayan
Doğum Yeri : Bingöl
Medeni Hali : Bekar
E-posta : acaglayan12@gmail.com
Adresi : Bingöl Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar
Mühendisliği Bölümü, 12000, Bingöl/Türkiye

Eğitim

Lise : Malatya Fen Lisesi (2000-2004)
Lisans : Hacettepe Üniversitesi Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü (2004 - 2009)
Yüksek Lisans : -
Doktora : Hacettepe Üniversitesi Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü (2011 - 2018)

Yabancı Dil ve Düzeyi

İngilizce : KPDS 79 (2010-Güz)

İş Deneyimi

Cybersoft Ltd. Şti. : Yazılım Mühendisi, 2010
Hacettepe Üniversitesi: Araştırma Görevlisi, 2011 – 2018
Bingöl Üniversitesi : Araştırma Görevlisi, 2018 – (Devam ediyor)

Deneyim Alanları

Bilgisayarlı Görü, Makine Öğrenmesi, Derin Öğrenme, Nesne Tanıma.

Tezden Üretilmiş Projeler ve Bütçeleri

-

Tezden Üretilmiş Yayınlar

- Ali Caglayan, Ahmet Burak Can, Exploiting Multi-Layer Features Using a CNN-RNN Approach for RGB-D Object Recognition, in Computer Vision – ECCV 2018 Workshops, **2018**.
- Ali Caglayan, Ahmet Burak Can, Volumetric Object Recognition Using 3-D CNNs on Depth Data, in IEEE Access, vol. 6, pp. 20058-20066, **2018**.
- Ali Caglayan, Ahmet Burak Can, An Empirical Analysis of Deep Feature Learning for RGB-D Object Recognition, in International Conference Image Analysis and Recognition, pp. 312-320. Springer, Cham, **2017**.
- Ali Caglayan, Ahmet Burak Can, 3D convolutional object recognition using volumetric representations of depth data, in Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on, pp. 125-128. IEEE, **2017**.

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

- 15th European Conference on Computer Vision (ECCV), Munich, Germany, 2018
- 14th International Conference on Image Analysis and Recognition (ICIAR), Montreal, Canada, 2017.
- 15th IAPR International Conference on Machine Vision Applications, Nagoya, Japan, 2017.



HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
~~YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU~~

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 06/12/2018

Tez Başlığı / Konusu: **Derin Öğrenme Tekniklerini Kullanarak RGB-D Nesne Tanıma**

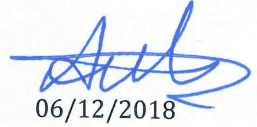
Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 118 sayfalık kısmına ilişkin, 06/12/2018 tarihinde ~~şahsım~~/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 2'dir.

Uygulanan filtrelemeler:

- 1- Kaynakça hariç
- 2- Alıntılar hariç
- 3- 5 kelimededen daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.


06/12/2018

Adı Soyadı: Ali ÇAĞLAYAN
Öğrenci No: N10268018
Anabilim Dalı: Bilgisayar Mühendisliği
Programı: Bilgisayar Mühendisliği
Statüsü: Y.Lisans Doktora Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.


Doç. Dr. Ahmet Burak CAN