Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Derleme Makalesi

Bilgisayarlı Görüde Öz-Denetimli Öğrenme Yöntemleri Üzerine Bir İnceleme

Serdar ALASU [a,*], Muhammed Fatih TALU [a]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İnönü Üniversitesi, Malatya, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: serdaralasu@gmail.com*
DOI: 10.29130/dubited.1201292

Öz

Derin öğrenme modelleri son on yılda görüntü sınıflandırma, nesne tespiti, görüntü bölütleme vb. bilgisayarlı görü görevlerinde büyük başarılar elde etmelerine rağmen denetimli öğrenme yaklaşımında olan bu modellerin eğitiminde büyük miktarda etiketli veriye ihtiyaç duyulmaktadır. Bu nedenle, son yıllarda insanlar tarafından manuel olarak etiketlenen veriye ihtiyaç duymadan etiketsiz büyük boyutlu veriden faydalanarak genelleştirilebilir görüntü temsillerini öğrenebilen öz-denetimli öğrenme yöntemlerine ilgi artmıştır. Bu çalışmada, bilgisayarla görü görevlerinde kullanılan öz denetimli öğrenme yöntemleri kapsamlı bir şekilde incelenmiş ve öz denetimli öğrenme yöntemlerinin kategorizasyonu sağlanmıştır. İncelenen öz-denetimli öğrenme yöntemlerinin görüntü sınıflandırma, nesne tespiti ve görüntü bölütleme hedef görevleri için performans karşılaştırmaları sunulmuştur. Son olarak, mevcut yöntemlerdeki sorunlu hususlar tartışılmakta ve gelecek çalışmalar için potansiyel araştırma konuları önerilmektedir.

Anahtar Kelimeler: *Bilgisayarlı Görü, Öz-Denetimli Öğrenme, Karşılaştırmalı Öğrenme*

A Review on Self-Supervised Learning Methods in Computer Vision

Abstract

Although deep learning models have achieved great success in computer vision tasks such as image classification, object detection, image segmentation in the last decade, a large amount of labeled data requires in the training of these models, which are in a supervised learning approach. Therefore, in recent years, there has been an increased interest in self-supervised learning methods that can learn generalizable image representations by utilizing large-scale unlabeled data without the need for manually labeled data by humans. In this study, self-supervised learning methods used in computer vision tasks are comprehensively reviewed and categorization of self-supervised learning methods is provided. Performance comparisons of the reviewed self-supervised learning methods for image classification, object detection and image segmentation target tasks are presented. Finally, problematic issues in current methods are discussed and potential research topics are suggested for future studies.

Keywords: *Computer Vision, Self-Supervised Learning, Contrastive Learning*

Geliş: 08/11/2022, Düzeltme: 06/04/2022, Kabul: 16/08/2023

I. GİRİŞ

Son on yılda, uzman bilgisine ihtiyaç duymadan otomatik olarak probleme özgü özniteliklerin çıkarılmasını sağlayan denetimli öğrenme yaklaşımına dayalı derin öğrenme modelleri, görüntü sınıflandırma [1]–[7], nesne tespiti [8]–[11], görüntü bölütleme [12]–[16] vb. bilgisayarlı görü görevlerinde yüksek performans elde etmişlerdir. Ancak denetimli öğrenme yaklaşımını kullanan derin öğrenme modelleri ayırt edici özniteliklerin çıkarılmasında, insanlar tarafından manuel olarak etiketlenen büyük miktarda veriye ihtiyaç duymaktadır [17]. Veri etiketlemeyi kolaylaştırmak için birçok etiketleme aracı [18] geliştirilmesine rağmen, etiketli verinin elde edilmesi emek yoğun, zaman alıcı ve pahalı bir işlemdir. Ayrıca medikal görüntü alanında görüntü elde etme zorluğu ve uzman bilgisi gerekliliği bu alanda etiketli veri elde etmeyi daha da zorlaştırmaktadır [19]. Tek bir insanın dakikada bir görüntü etiketlediği ve etiketleme dışında başka bir iş yapmadığı bir durumda 14 milyondan fazla görüntü ve 20 binden fazla sınıftan oluşan ImageNet [20] veri setini etiketlemesi için 26 yıldan fazla zaman gerekmektedir.

Etiketli verinin az olduğu durumlar için genel yaklaşım denetimli öğrenime dayalı transfer öğrenimidir. Transfer öğrenimi yaklaşımında, kaynak görevden öğrenilen bilgi hedef göreve aktarılmaktadır [21]. Denetimli öğrenime dayalı transfer öğrenimi, genelleştirilebilir temsillerin öğrenilebilmesi için derin öğrenme modelinin etiketli çok büyük kaynak veri setiyle eğitilmesi ve bu temsillerin az sayıda etiketli veriden oluşan hedef görevde kullanılması aşamalarından oluşmaktadır [22]. ImageNet [20] veri setiyle eğitilmiş, AlexNet [3], VGG16 [4], ResNet [5], DenseNet [7], GoogleNet [6], InceptionV3 [1] ve EfficientNet [2] derin evrişimsel sinir ağı(DESA) mimarilerinin ön-eğitimli derin öğrenme modelleri birçok uygulamada yaygın olarak kullanılmaktadır [23]. Ancak denetimli öğrenme temelli transfer öğrenimi yaklaşımında, ön-eğitim aşamasında çok büyük etiketli veriye ihtiyaç duyulmakta ve kaynak veri setinin hedef veri setinden çok farklı olduğu durumlarda kaynak görevden elde edilen temsillerin hedef görev için genelleştirilmesi problemi yaşanmaktadır [24].

Denetimli öğrenme temelli transfer öğrenimindeki bu sorunlar nedeniyle, son yıllarda etiketsiz veriden otomatik olarak elde edilen etiketlerle derin öğrenme modelleri eğitilerek genelleştirilebilir temsiller elde edebilen öz-denetimli öğrenme yaklaşımları ortaya çıkmıştır. İnternetteki etiketsiz büyük boyutlu veriden faydalanan öz-denetimli öğrenme yöntemleri denetimli öğrenme yöntemleriyle rekabet edecek sonuçlar elde etmişlerdir [25]–[30]. Öz-denetimli öğrenmede, denetim sinyali etiketsiz veriden otomatik olarak oluşturulmakta ve yardımcı görev olarak adlandırılan görevin çözümünde bu denetim sinyali kullanılarak derin öğrenme modeli eğitilmektedir. Yardımcı görev ile elde edilen ön-eğitimli ağ, az sayıda etiketli veriye sahip görüntü sınıflandırma, nesne tespiti ve görüntü tespiti vb. hedef görevlerde kullanılmaktadır. Böylece hedef görevde kullanılacak derin öğrenme modelinin ağırlıklarının rastgele değerlerle başlatılarak sıfırdan öğrenilmesi yerine, yardımcı görevle öğrenilen ağırlıklarla derin öğrenme modelinin eğitilmesi sağlanmaktadır.

Bu çalışmada bilgisayarlı görü görevlerinde kullanılan öz denetimli öğrenme yöntemlerinin kapsamlı incelemesi yapılarak öz denetimli öğrenme yöntemleri kategorize edilmiştir. İncelenen öz-denetimli öğrenme yöntemlerinin farklı veri setlerindeki ve görüntü sınıflandırma, nesne tespiti ve görüntü bölütleme hedef görevlerindeki performans karşılaştırmaları sunulmuştur. Son olarak, mevcut yöntemlerdeki sorunlu hususlar tartışılmakta ve gelecek çalışmalar için potansiyel araştırma konuları önerilmektedir.

Makalenin geri kalanı şu şekilde düzenlenmiştir: II. Bölümde, öz-denetimli öğrenme yöntemleri kategorize edilerek, öz-denetimli öğrenme yöntemleri detaylı olarak anlatılmıştır. III. Bölümde öz-denetimli öğrenme yöntemlerinin farklı veri setlerindeki ve farklı bilgisayarlı görü görevlerindeki performansları sunulmuştur. IV. Bölümde tartışma ve V. Bölümde ise sonuç kısmı yer almaktadır.

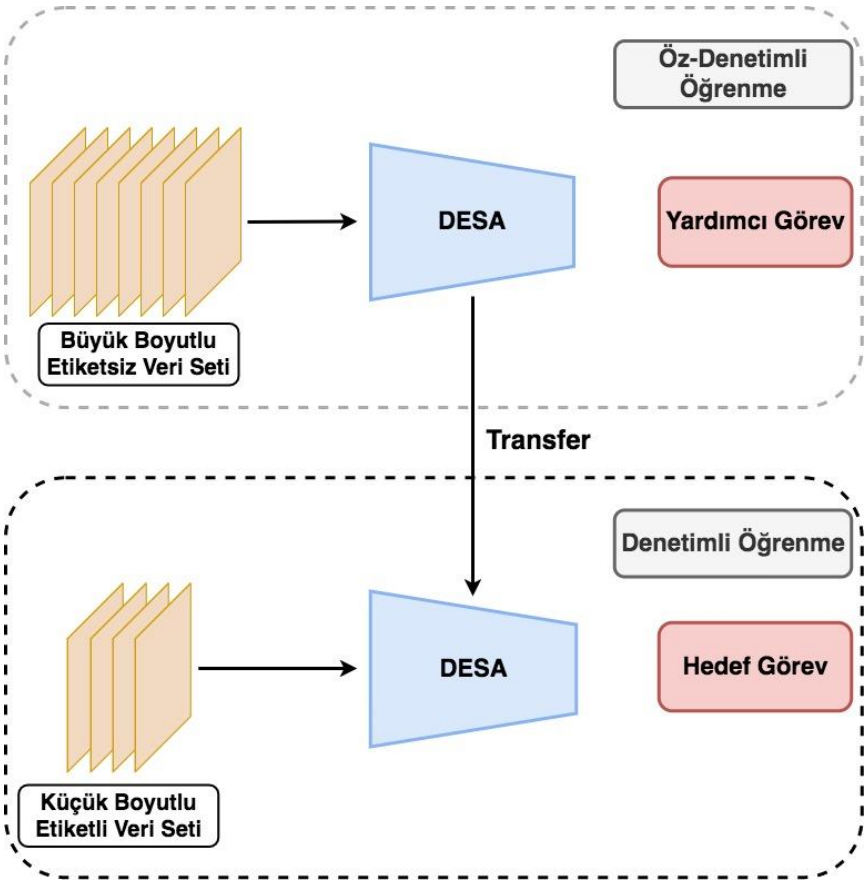
II. ÖZ-DENETİMLİ ÖĞRENME

Etiketli veri elde etmenin zaman alıcı, pahalı ve emek yoğun bir işlem olması ve internetteki etiketsiz büyük boyutlu verinin varlığı, araştırmacıların öz-denetimli öğrenme yaklaşımına ilgi duymalarına neden olmuştur [31]. Öz-denetimli öğrenmede, verilerin insanlar tarafından manuel olarak etiketlenmesine gerek duyulmadan, verinin kendi özellikleri kullanılarak çok büyük boyutlu veri otomatik olarak etiketlenmekte ve bu etiketler sözde etiket olarak isimlendirilmektedir. Öz-denetimli öğrenme yaklaşımı yardımcı görev ve hedef görevleri içermektedir. Gerçekte çözmek istediğimiz görüntü sınıflandırma, nesne tespiti, görüntü bölütleme vb. görevler hedef görev olarak adlandırılırken, hedef görevleri çözmeye yardımcı olan görevler yardımcı görev olarak adlandırılmaktadır [32]. Elde edilen etiketli veri, öz-denetimli öğrenme için tasarlanan yardımcı görevde eğitilerek genelleştirilebilir görüntü temsili öğrenilmektedir [33], [34]. DESA mimarilerinin çözmesi istenilen yardımcı görev ve otomatik olarak sahte etiketlerin elde edilmesi öz-denetimli öğrenmenin ön-eğitim aşamalarını oluşturmaktadır. Ön-eğitim aşamasında elde edilen genelleştirilebilir görüntü temsili, az sayıda etiketli verinin olduğu hedef göreve aktarılmaktadır. Genelleştirilebilir temsil ile hedef görevlere kolayca uyarlanabilir olması kastedilmektedir [35]. Böylece hedef görevde model ağırlıklarının rastgele atanmış değerlerle başlatılması yerine yardımcı görevde elde edilen ön-eğitimli ağ kullanılmaktadır.

Öz-Denetimli Öğrenme aşamaları:

1. Çok büyük boyutlu etiketsiz veri setinde otomatik olarak etiketlerin oluşturulması
2. Oluşturulan sözde etiketler ile yardımcı görevde eğitilerek genelleştirilebilir görüntü temsillerin öğrenilmesi
3. Elde edilen görüntü temsillerin, az sayıda etiketli veriye sahip hedef göreve aktarılması

Şekil 1'de öz-denetimli öğrenme yaklaşımının genel mimarisi gösterilmiştir.



Şekil 1. *Öz-denetimli öğrenme yaklaşımının genel mimarisi*

Denetimli öğrenme yaklaşımı görsel temsillerin çıkarılmasında insanlar tarafından manuel olarak oluşturulan etiketli veriye ihtiyaç duyarken, denetimsiz öğrenmede herhangi bir etiket bilgisine ihtiyaç duyulmamaktadır. Öz-denetimli öğrenmede ise insanlar tarafından manuel oluşturulan etiketli veri kullanılmadığından denetimsiz öğrenmenin bir alt alanı olarak görülmektedir [31], [36]. Ancak denetimsiz öğrenmede öğrenme sürecinde herhangi bir etiket bilgisine ihtiyaç duyulmazken, öz-denetimli öğrenmede verinin kendisinden otomatik olarak üretilen sahte etiketler yardımcı görevde eğitilerek öğrenme gerçekleştirilmektedir [32].

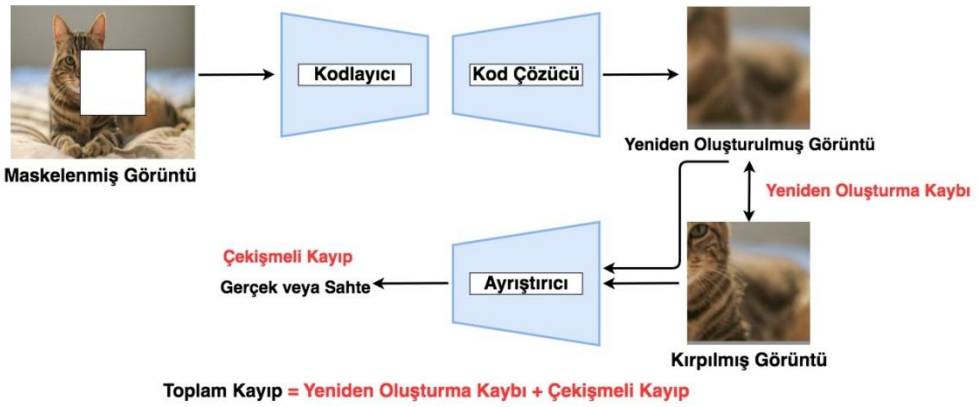
Öz-denetimli öğrenme yaklaşımını görüntü temsilinde kullanılan yardımcı göreve göre beş kategoriye ayrılabilir:

1. Yeniden oluşturma temelli öz-denetimli öğrenme
2. Tahmin temelli öz-denetimli öğrenme
3. Kümeleme temelli öz-denetimli öğrenme
4. Karşılaştırmalı öz-denetimli öğrenme
5. Karşılaştırmalı olmayan öz-denetimli öğrenme

A. YENİDEN OLUŞTURMA TEMELLİ ÖZ-DENETİMLİ ÖĞRENME

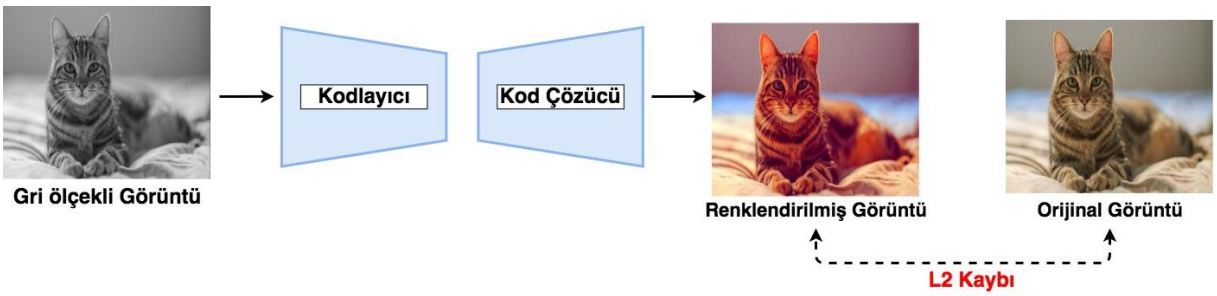
Yeniden oluşturma temelli öz-denetimli öğrenmede, orijinal görüntü üzerinde değişiklik yapılarak yeni bir görüntü oluşturulmakta ve daha sonra değişiklik yapılmış görüntüyü orijinal görüntüye eşleyen bir oto kodlayıcı eğitilmektedir. Kodlayıcı-kod çözücü mimarisinin kullanıldığı oto kodlayıcılar [35], [37] yeniden oluşturma temelli öz-denetimli öğrenmenin ilk örnekleridir. Kodlayıcı ağı girdi görüntüsünü temsil vektörüne indirgerken, kod çözücü ağ ise temsil vektöründen orijinal görüntüyü yeniden oluşturmaya çalışmaktadır.

Yeniden oluşturma temelli öz-denetimli öğrenme yaklaşımının kullanım alanlarından biri maskelenmiş bir görüntü bölgesinin tahmin edilmesi problemidir. Görüntü maskeleme [38] çalışmasında görüntünün maskelenen kısmı, diğer kısımlarından faydalanılarak yeniden oluşturulmaktadır. Bu işlem ağın eğitilmesi için bir yardımcı görev olarak kullanılmaktadır. Kodlayıcı-kod çözücü ağ mimarisine sahip bir DESA, maskelenmiş bölgeleri olan çok sayıda etiketlenmemiş görüntü kullanılarak eğitilmektedir. Maskelenmiş görüntüyü girdi olarak alan kodlayıcı görüntü temsilini öğrenirken, kod çözücü görüntü temsilini kullanarak maskelenmiş bölgeyi yeniden oluşturmaktadır. Model, yeniden oluşturma kaybı ve çekişmeli kaybın birlikte kullanılmasıyla eğitilmektedir. Yeniden oluşturma kaybı (*L2*), görüntünün eksik bölgesinin genel yapısının öğrenilmesini sağlarken, çekişmeli kayıp ise yeniden oluşturulan maske görüntüsünün daha gerçekçi olmasını sağlamaktadır. Bu yardımcı görevi çözebilmek için, model görüntüdeki farklı nesnelerin yapısını, renklerini ve görüntünün anlamsal özniteliklerini öğrenmek zorundadır. Şekil 2'de görüntü maskeleme problemi için yeniden oluşturma temelli öz-denetimli öğrenme yaklaşımının kullanımı gösterilmektedir.



Şekil 2. *Görüntü maskeleme [38] problemi için yeniden oluşturma temelli öz-denetimli öğrenme yaklaşımı*

Yeniden oluşturma temelli öz-denetimli öğrenme yaklaşımının uygulandığı diğer bir problem görüntü renklendirmedir [39]. Bu problemde, girdi görüntüsü olarak gri ölçekli bir görüntü kullanılmakta ve bu görüntünün renkli versiyonu yeniden oluşturulmaya çalışılmaktadır. *Lab* renk uzayı kullanılan çalışmada, modele girdi olarak görüntünün *L* kanalı verilmekte ve model görüntünün *a* ve *b* renk kanallarını tahmin etmeye çalışmaktadır. Her pikselin doğru şekilde renklendirilebilmesi için modelin nesneleri tanıması ve ilişkili kısımların piksellerini birlikte gruplandırması gerekmektedir. Böylece model görsel temsilleri öğrenebilmektedir. Bu çalışmada, kod çözücü ağında temsilleri öğrenen kodlayıcı kısmında ise renklendirilmeyi gerçekleştiren tamamıyla evrişimsel sinir ağı kullanılmıştır. Model, tahmin edilen renk ve orijinal renk arasındaki yeniden oluşturma (*L2*) kaybıyla eğitilmektedir. Yardımcı problem olarak görüntü renklendirme probleminin kullanımı Şekil 3'te gösterilmiştir.

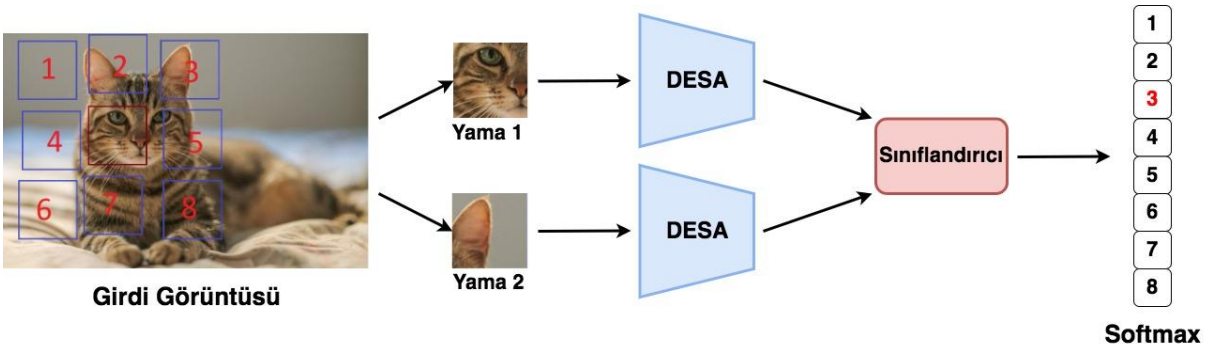


Şekil 3. *Görüntü renklendirme [39] problemi için yeniden oluşturma temelli öz-denetimli öğrenme yaklaşımı*

B. TAHMİN TEMELLİ ÖZ-DENETİMLİ ÖĞRENME

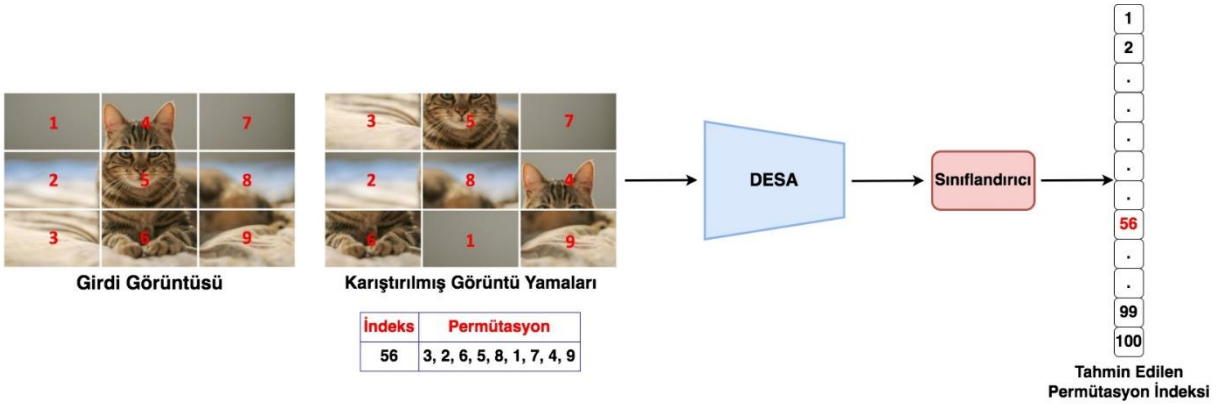
Tahmin temelli öz-denetimli öğrenmede, verilerin zengin uzamsal bilgisinden yararlanılarak etiketli veri otomatik olarak oluşturulmaktadır. Tahmin temelli yardımcı görevlerin kullanıldığı öz-denetimli öğrenme yaklaşımlarında, bir DESA otomatik olarak oluşturulan etiketli veriyi tahmin etmeye çalışarak eğitilmekte ve genelleştirilebilir görüntü temsili elde edilmeye çalışılmaktadır.

Tahmin temelli öz-denetimli öğrenme yaklaşımının uygulandığı problemlerden ilki görece pozisyon tahminidir [40]. Bu problemde amaç, rastgele seçilen yamanın merkez yamaya göre pozisyonunu tahmin etmektir. Bunun için ilk olarak görüntüler 3x3 büyüklüğünde yamalara ayrılmaktadır. Daha sonra biri merkez yama, diğeri merkeze komşu 8 yamadan olmak üzere rastgele yama çiftleri oluşturulmaktadır. Yama çiftleri, AlexNet [3] mimarisinin omurga olarak kullanıldığı model ağırlıklarını paylaşan ikiz DESA'ya girdi olarak verilmekte ve bu iki DESA'dan elde edilen temsiller birleştirilmektedir. Birleştirilen temsiller kullanılarak rastgele seçilen yamanın merkez yamaya göre pozisyonu tahmin edilmektedir. Görüntüler yamalara ayrılırken yamalar arasında boşluk bırakılıp yama pozisyonlarında değişiklikler yapılarak hem görevin karmaşıklığı artırılmakta hem de DESA'nın basit çözümlere ulaşması engellenmektedir. Şekil 4'te görece pozisyon tahmini problemi için tahmin temelli öz-denetimli öğrenme yaklaşımının kullanımı gösterilmektedir.



Şekil 4. *Görece pozisyon tahmini [40] problemi için tahmin temelli öz-denetimli öğrenme yaklaşımı*

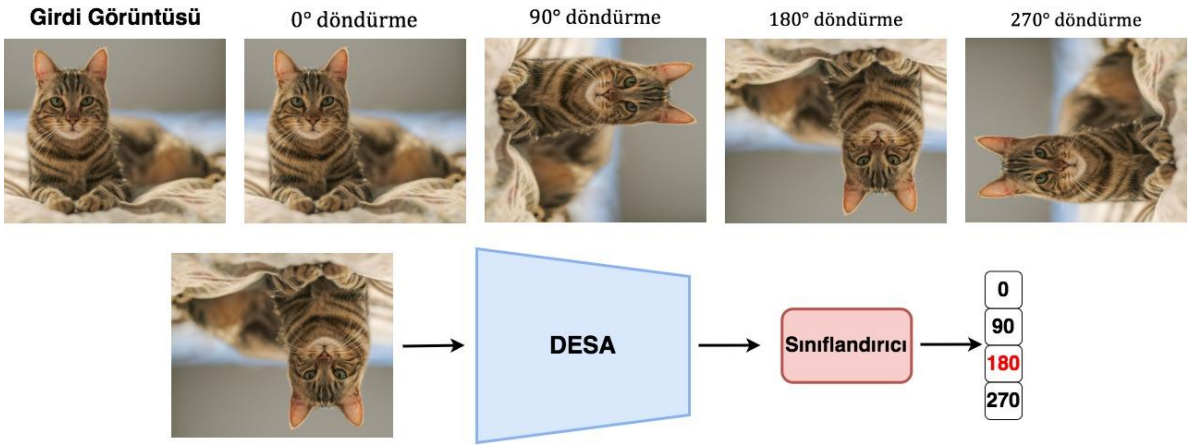
Tahmin temelli öz-denetimli öğrenme yaklaşımının kullanıldığı diğeri bir problem yapboz bulmacadır [41]. Yardımcı görev olarak yapboz bulmaca çözümünün önerildiği bu çalışmada, ilk olarak görüntü 3x3 büyüklüğünde yamalara ayrılmakta, sonrasında ise bu yamalar rastgele bir şekilde karıştırılarak yapboz bulmaca oluşturulmaktadır. Bu yamaların orijinal yerlerini bulacak şekilde DESA eğitilerek, DESA'nın nesneleri tanımayı ve nesneleri oluşturan parçaların uzamsal ilişkisini öğrenmesi istenmektedir. Görece pozisyon tahmini [40] çalışmasına benzer şekilde görüntüler yamalara ayrılırken yamalar arasında rastgele büyüklükte boşluklar bırakılarak DESA'nın basit çözümlere ulaşması engellenmektedir. Ayrıca, 9 yama için 362.880 (9!) olası permütasyon çok büyük bir çözüm uzayına neden olduğundan, bu kadar büyük bir çözüm uzayından kaçınmak için, DESA'nın eğitiminde en yüksek hamming mesafesine sahip belirli sayıdaki permütasyon seçilerek olası permütasyonların bir alt kümesi kullanılmaktadır. Bu çalışmada yazarların Bağlamdan Bağımsız Ağ olarak adlandırdığı DESA kullanılmaktadır. Ağırlıkların paylaşıldığı Bağlam Bağımsız Ağ'da 9 yama girdi olarak kabul edilmekte ve ağın çıktıları birleştirilerek permütasyon kümesi tahmin edilmeye çalışılmaktadır. Şekil 5'te yapboz bulmaca çözme problemi için tahmin temelli öz-denetimli öğrenme yaklaşımının kullanımı gösterilmiştir.



İndeks	Permütasyon
56	3, 2, 6, 5, 8, 1, 7, 4, 9

Şekil 5. *Yapboz bulmaca çözme [41] problemi için tahmin temelli öz-denetimli öğrenme yaklaşımı*

Bir diğer tahmin temelli öz-denetimli öğrenme yönteminde yardımcı görev olarak bir görüntüye uygulanan döndürülme açısının tahmini kullanılmaktadır [42]. Görüntülerin 0°, 90°, 180°, 270° açılar ile döndürülmesiyle elde edilen görüntüler ve görüntülerin döndürülme açıları etiketli veri setini oluşturmaktadır. Bu veri seti oluşturulduktan sonra, dört sınıflı bir sınıflandırma görevine benzer şekilde, bir DESA giriş görüntüsünün döndürüldüğü açının tahminiyle eğitilmektedir. Doğru tahmin için görüntüdeki nesnelerin yüksek seviyede anlaşılması gerekmektedir. Şekil 6'da görüntü döndürme açısı tahmin probleminin öz-denetimli çözümü gösterilmektedir.

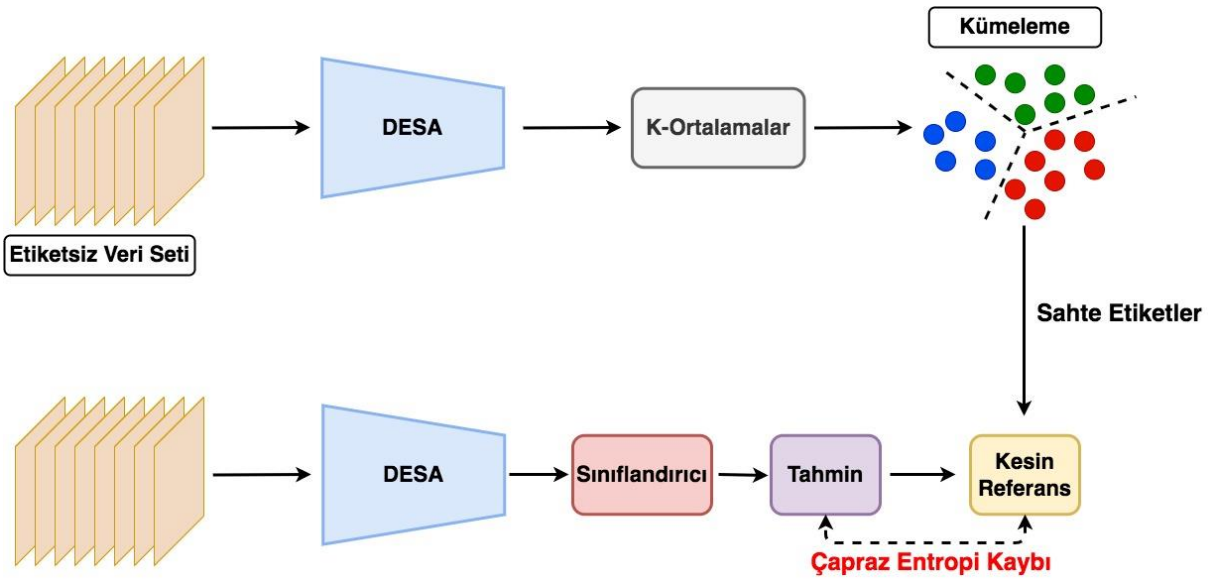


Şekil 6. *Döndürme açısı tahmini probleminin öz-denetimli çözümü [42]*

C. KÜMELEME TEMELLİ ÖZ-DENETİMLİ ÖĞRENME

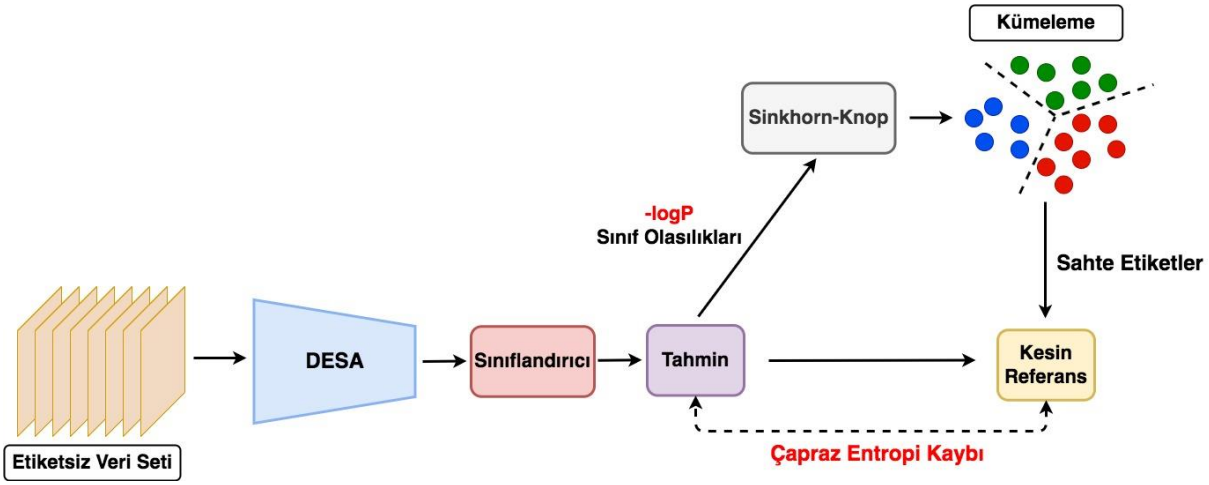
Kümeleme temelli öz-denetimli öğrenme yöntemlerinde, DESA'nın eğitiminde kullanılacak etiketler, görüntü temsillerinin kümelenmesiyle otomatik olarak elde edilmektedir. Kümeleme temelli öz-denetimli öğrenme yöntemleri çevrim dışı ve çevrim içi yöntemler olarak ikiye ayrılmaktadır. Çevrim dışı yöntemlerde, veri kümesindeki tüm görüntüler DESA'dan en az bir defa geçirilerek elde edilen temsiller kümelenmektedir. Bu nedenle büyük boyutlu veri setleri için çevrim dışı yöntemler yüksek hesaplama gücüne ihtiyaç duymaktadırlar. Çevrim içi yöntemlerde ise kümeleme ve öğrenme eş zamanlı olarak gerçekleştirilmektedir.

Kümelemeyi öz-denetimli öğrenmede kullanan ilk yöntemlerden biri derin kümeleme [43] çalışmasıdır. Bu çalışmada, birbirini takip eden kümeleme ve öğrenme aşamalarından oluşan bir çevrim dışı kümeleme temelli öz-denetimli öğrenme yaklaşımı kullanılmaktadır. Derin kümeleme [43] yönteminde, görüntüler bir DESA'dan geçirilmekte ve elde edilen temsillere temel bileşenler analizi, beyazlatma ve L2 normalizasyon uygulanmaktadır. Boyutu indirgenen ve normalize edilen temsiller daha sonra k-ortalamalar yöntemiyle kümelenerek sözde etiketler oluşturulmakta ve bu etiketler kullanılarak DESA'nın eğitimi gerçekleştirilmektedir. Bu iki aşama tekrarlı bir şekilde çalıştırılarak genelleştirilebilir görüntü temsilleri elde edilmeye çalışılmaktadır. Şekil 7'de derin kümeleme [43] yaklaşımı gösterilmektedir.



Şekil 7. *Çevrim dışı kümeleme temelli öz-denetimli öğrenme: Derin Kümeleme [43]*

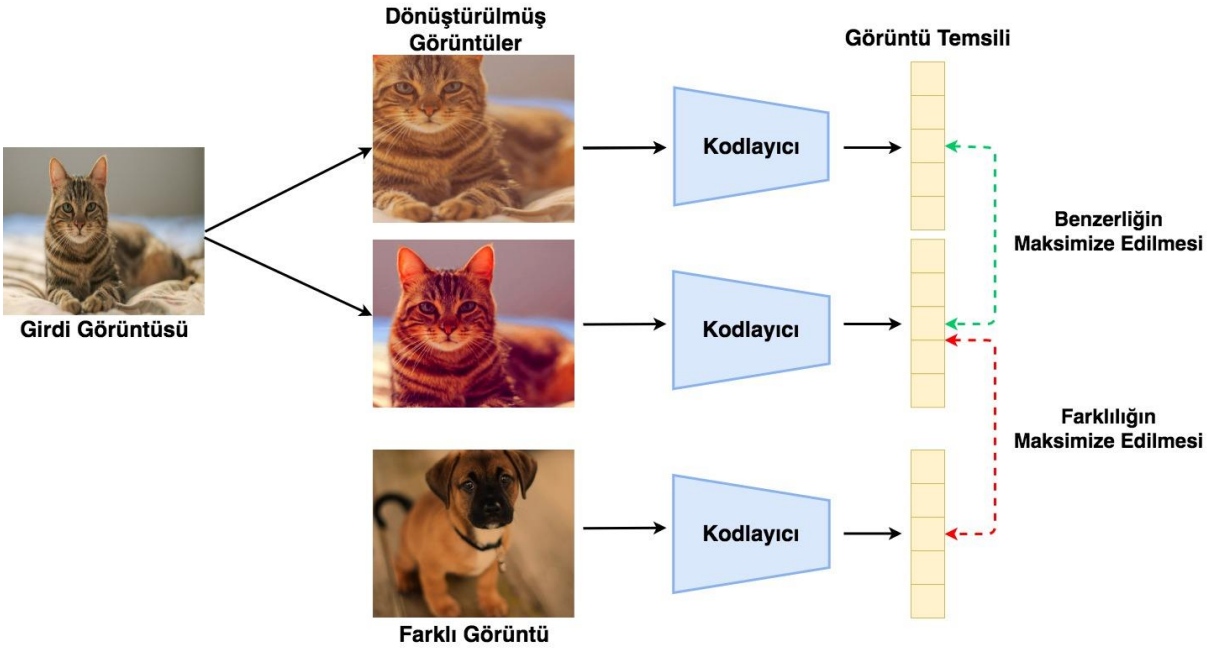
Kümeleme ve öğrenmenin eş zamanlı olarak gerçekleştirildiği SeLA [44], çevrim içi kümeleme temelli öz-denetimli öğrenme yöntemidir. Bu yöntemde kümeleme ve öğrenme için tek bir amaç fonksiyonu kullanılırken, kümeleme en uygun aktarım problemlerinin çözümünde kullanılan Sinkhorn-Knopp algoritmasıyla [45] gerçekleştirilmektedir. Sinkhorn-Knopp algoritması [45], DESA'nın tahmininden elde edilen sınıf olasılıklarının oluşturduğu maliyet matrisi ve her kümeye eşit sayıda görüntü atanması kısıtını kullanarak kümelemeyi gerçekleştirmektedir. Şekil 8'de SeLA [44] yönteminin gösterimi yapılmıştır.



Şekil 8. *Çevrim içi kümeleme temelli öz-denetimli öğrenme: SeLA [44]*

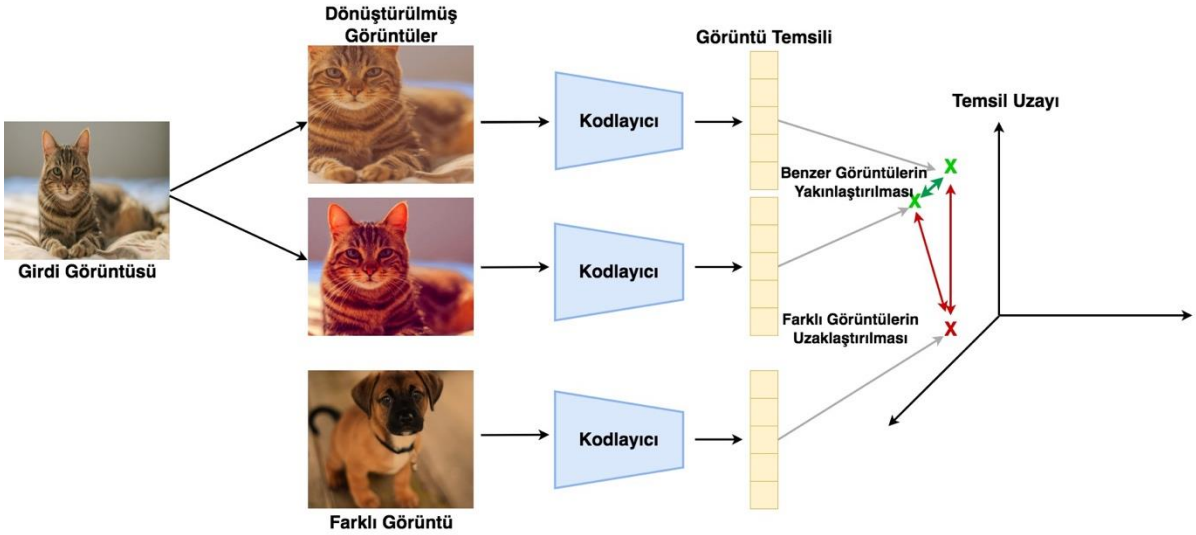
D. KARŞILAŞTIRMALI ÖZ-DENETİMLİ ÖĞRENME

Son yıllarda yapılan çalışmalarda, karşılaştırmalı öz-denetimli öğrenmeyle ön-eğitilen DESA'ların denetimli öğrenmeyle ön-eğitilen DESA'lar ile rekabet edecek kabiliyette olduğu görülmüştür [25], [26], [30]. Bu yaklaşımda, DESA, benzer görüntü çiftlerini farklı görüntü çiftlerinden ayırt edebilecek genelleştirilebilir görüntü temsillerini üretmek için eğitilmektedir. Görüntü benzerliği, görüntülerin farklı dönüşümlerinin kullanımıyla otomatik bir şekilde tanımlanmaktadır. Aynı görüntüden farklı görüntü dönüşümleriyle oluşturulan birbirine benzer görüntü çiftleri *pozitif çift* ve farklı görüntülerin dönüşümlerinin oluşturduğu birbirinden farklı görüntü çiftleri ise *negatif çift* olarak adlandırılmaktadır. Karşılaştırmalı öğrenmeye benzer şekilde, insan beyni bir nesneyi diğerinden ayırt etmek için nesnenin tüm bilgisine ihtiyaç duymadan sadece ayırt edici özelliklerini tutmaktadır [46]. Şekil 9'da karşılaştırmalı öz-denetimli öğrenme yaklaşımının genel gösterimi yapılmıştır.



Şekil 9. *Karşılaştırmalı öz-denetimli öğrenme yaklaşımının genel gösterimi*

Karşılaştırma temelli öz-denetimli öğrenme yöntemlerinde DESA'nın eğitiminde karşılaştırmalı kayıp fonksiyonu kullanılmaktadır. Karşılaştırmalı kayıp fonksiyonu, temsil uzayında aynı görüntünün farklı dönüşümlerinden oluşturulan pozitif çiftlerin temsillerini birbirine yaklaştırmaya zorlarken, farklı görüntü çiftlerinin dönüşümlerinden oluşturulan negatif çiftlerin temsillerini ise birbirinden uzaklaştırmaya zorlamaktadır. Şekil 10'da karşılaştırmalı kayıp fonksiyonun çalışma prensibi gösterilmiştir.



Şekil 10. *Karşılaştırmalı kayıp fonksiyonun çalışma prensibinin gösterimi*

x orijinal görüntüyü, x^+, x orijinal görüntüsüne benzeyen pozitif örneği, x^-, x orijinal görüntüsünden farklı negatif örneği, sim() görüntü benzerlik fonksiyonunu göstermek üzere, karşılaştırmalı öğrenmede amaç $sim\big(f(x), f(x^+)\big) \gg sim\big(f(x), f(x^-)\big)$ şartını sağlayan f kodlayıcısını öğrenmektir. Bu amacı gerçekleştirmek için karşılaştırmalı kayıp fonksiyonu kullanılmaktadır. Denetimli öğrenme uygulamalarında [47]–[49] kullanılmaya başlayan karşılaştırmalı kayıp fonksiyonu, öz-denetimli öğrenmede ilk olarak InstDisc [50] çalışmasında kullanılmıştır. Her bir

görüntü örneğinin ayrı bir sınıf olarak ele alındığı InstDisc [50] çalışmasında, DESA sınıfları ayırt etmek için eğitilmektedir. Her görüntünün ayrı bir sınıf olarak kullanılması softmax fonksiyonun hesaplanmasında zorluk yaratacağı için, bu çalışmada softmax fonksiyonuna yakınsayabilen çok sınıflı sınıflandırma problemini ikili sınıflandırma problemine dönüştüren NCE [51] kayıp fonksiyonu kullanılmaktadır.

NCE [51] kayıp fonksiyonun çok sayıda negatif çift içeren bir varyantı olan InfoNCE [52] karşılaştırmalı öz-denetimli öğrenme yöntemlerinde sıklıkla kullanılan diğer bir kayıp fonksiyonudur. z_i , z^+ aynı görüntünün çoğaltılmış görüntülerinin temsilleri, z_i , z^- farklı görüntülerin çoğaltılmış görüntülerinin temsilleri, τ sıcaklık derecesi (softmax sonucu elde edilen olasılık dağılımındaki rastgeleliğin kontrolü için kullanılan bir parametre) ve $N-1$ negatif çift sayısını göstermek üzere, Eşt. 1'de InfoNCE [52] kayıp fonksiyonu verilmiştir.

$$\mathcal{L}_{InfoNCE} = -log \frac{\exp(sim(z_i, z^+)/\tau)}{\exp(sim(z_i, z^+)/\tau) + \sum_{j=1}^{N-1} \exp(sim(z_i, z_j^-)/\tau)} \quad (1)$$

Görüntü çiftlerinin temsillerinin arasındaki yakınlığın ölçümünde benzerlik metrikleri kullanılmaktadır. İki vektör arasındaki kosinüs açısını hesaplayan kosinüs benzerliği görüntü temsillerinin benzerliğinin hesaplanmasında sıklıkla kullanılmaktadır. Çoğaltılmış görüntülerin temsilleri z_i ve z_j olmak üzere, görüntü temsilleri arasındaki kosinüs benzerliği $s_{i,j}$ Eşt. 2'de verilmiştir.

$$s_{i,j} = \frac{z_i^T z_j}{(\| z_i \| \| z_j \|)} \quad (2)$$

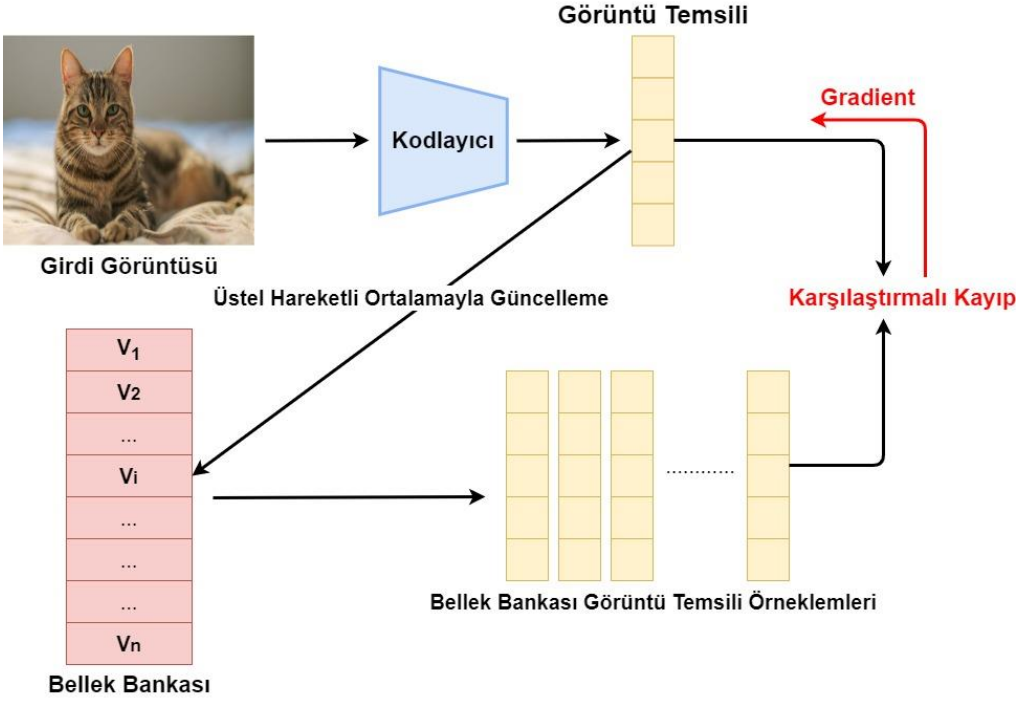
InfoNCE [52] kayıp fonksiyonu karşılıklı bilgi tahmincisi olarak da kullanılmaktadır. Pozitif çiftin görüntülerinin temsilleri arasındaki karşılıklı bilgiyi maksimize ederek genelleştirilebilir temsil öğrenilmeye çalışılmaktadır. Ancak yüksek boyutlu rastgele değişkenler arasındaki karşılıklı bilginin hesaplanma zorluğu nedeniyle InfoNCE [52] amaç fonksiyonu kullanılarak rastgele değişkenler arasındaki karşılıklı bilginin alt sınırı tahmin edilmektedir. Eşt. 3'te de görüleceği üzere pozitif çifti oluşturan görüntü temsillerinin arasındaki karşılıklı bilginin alt sınırının maksimize edilebilmesi için InfoNCE [52] kayıp fonksiyonunun minimize edilmesine ve negatif çift sayısının artırılmasına ihtiyaç duyulmaktadır. Bu nedenle karşılaştırmalı öğrenmede genelleştirilebilir görüntü temsilinin elde edilmesinde çok sayıda negatif çifte ihtiyaç duyulmaktadır.

$$I(z_i, z^+) \geq \log(N) - \mathcal{L}_{InfoNCE} \quad (3)$$

Karşılaştırma temelli öz-denetimli öğrenme yöntemlerinin ilklerinden olan Exemplar [53] yönteminde, veri setinden rastgele seçilen örnek için görüntü sınıfı oluşturmakta ve bir DESA bu örnekleri ayırt edecek şekilde eğitilmektedir. Bu çalışmada etiketsiz veri setinden rastgele seçilen görüntülerden 32x32 büyüklüğünde kırpılan görüntü yamalarına anlamsal olarak bir değişikliğe neden olmayacak şekilde çok sayıda rastgele görüntü dönüşümleri uygulanarak elde edilen görüntülerle vekil sınıflar oluşturulmaktadır. Böylece etiketsiz veri setindeki her bir görüntü için bir sınıf olacak şekilde etiketli bir veri seti oluşturulmaktadır. Bir DESA, oluşturulan vekil sınıfları ayırt edebilmek için eğitilerek, genelleştirilebilir görüntü temsilleri elde edilmeye çalışılmaktadır. Şekil 11'de seçilen görüntü yaması ve bu görüntü yamasına uygulanan rastgele dönüşümlerle oluşturulan vekil sınıfa ait görüntüler gösterilmiştir.

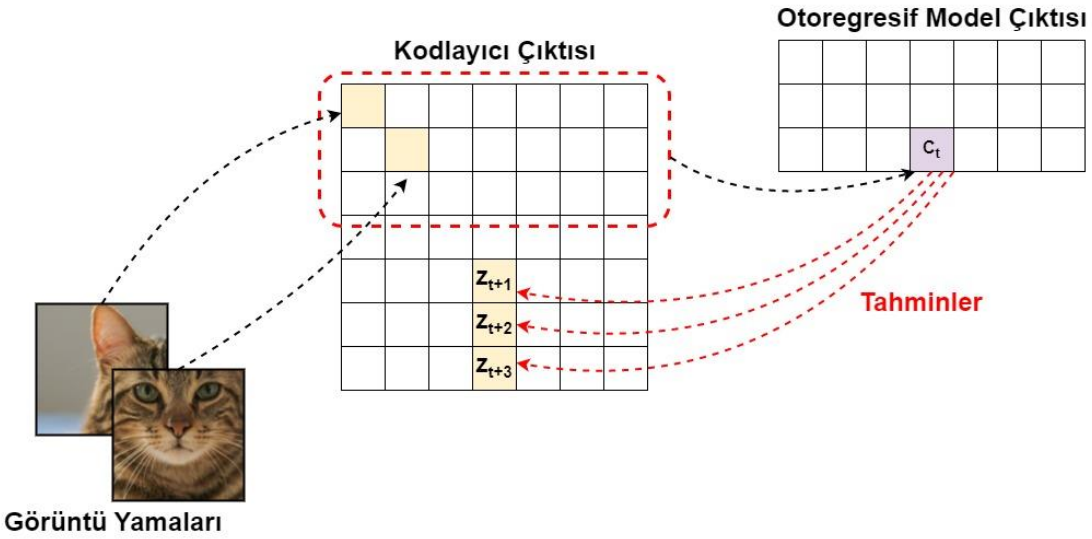
Şekil 11. *Exemplar [53] yöntemi için oluşturulmuş vekil sınıf örneği*

Exemplar [53] çalışmasında veri setinden rastgele seçilen görüntüler için bir vekil sınıfın oluşturulması seçilen görüntü sayısı arttıkça softmax fonksiyonun hesaplanmasında zorluk yaşanmasına neden olmaktadır. InstDisc [50] çalışması softmax fonksiyonuna yakınsayabilen çok sınıflı sınıflandırma problemini ikili sınıflandırma problemine dönüştüren NCE [51] karşılaştırmalı kayıp fonksiyonunu kullanarak Exemplar [53] yöntemindeki soruna çözüm getirmiştir. Bu çalışmada, her bir görüntü örneği ayrı bir sınıf olarak ele alınmaktadır ve DESA her bir görüntü örneğinin sınıflarını ayırt etmek için eğitilmektedir. Böylece görsel olarak benzer görüntülerin birbirine yakın ve görsel olarak farklı görüntülerin birbirinden uzak olarak haritalandığı görüntü temsil uzayı elde edilmeye çalışılmaktadır. Parametrik softmax fonksiyonunda bir sınıf prototipi olarak hizmet eden ağırlık vektörleri kullanılırken, bu çalışmada görüntü temsillerinin doğrudan karşılaştırıldığı parametrik-olmayan bir softmax fonksiyonu önerilmektedir. Ayrıca bu çalışmada görüntü temsilleri her iterasyonda yeniden hesaplanmak yerine bellek bankasında tutulmakta ve bellek bankasındaki temsiller her iterasyonda üstel hareketli ortalamayla güncellenmektedir. Şekil 12'de InstDisc [50] yöntemi gösterilmiştir.



Şekil 12. *InstDisc [50] yönteminin gösterimi*

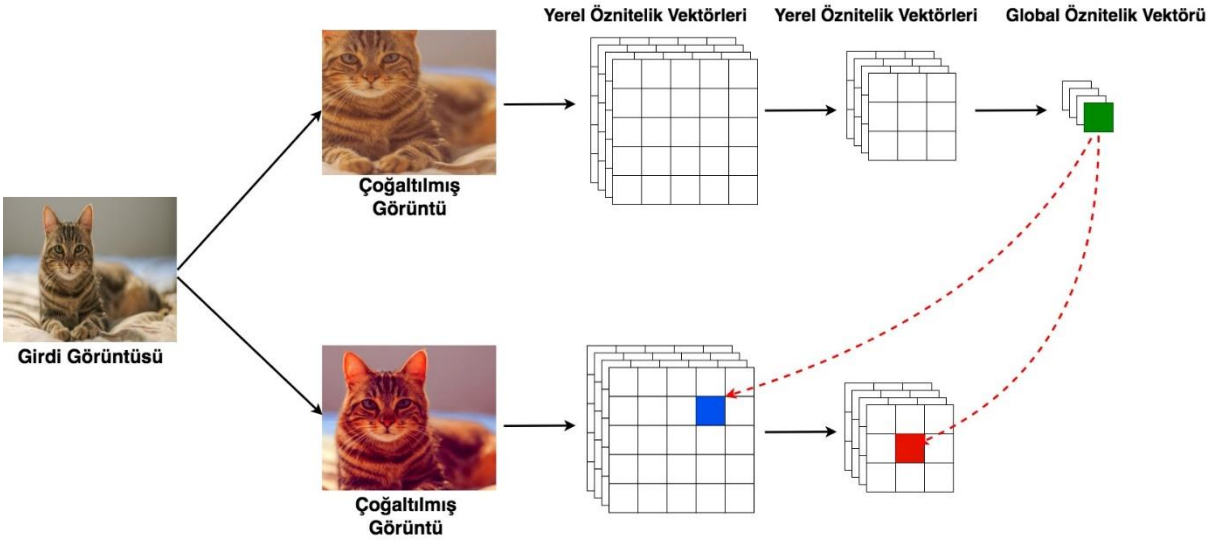
Karşılaştırmalı öğrenme yöntemlerinden bir diğeri olan CPC [52], ardışık veri olarak modellenebilen ses, metin, görüntü, video vb. farklı veri tiplerine uygulanabilen bir yöntemdir. Bu çalışmada görüntü temsilinin öğrenilmesinde, bir sistemin mevcut ve geçmiş durumlardan faydalanılarak gelecekteki durumları tahmin etmeye çalışan tahmine dayalı kodlama ve karşılaştırmalı öğrenme beraber kullanılmaktadır. CPC [52] yönteminde, bir görüntü ardışık görüntü yamaları olarak görülmekte ve görüntüler için zaman çizelgesi görüntünün sol üstü geçmiş ve sağ altıda gelecek olacak şekilde ele alınmaktadır. Bu yöntemde ilk olarak, 256x256 boyutundaki giriş görüntüsünden 64x64 boyutunda ve komşu yamaların 32 pikseli üst üste gelecek şekilde yamalar kırpılmakta ve bu yamalar bir kodlayıcıdan geçirilerek bu yamalara ait temsiller elde edilmektedir. Yamalara ait görüntü temsilleri elde edildikten sonra, PixelCNN [54] otoregresif modeliyle geçmişteki yama temsillerinden faydalanılarak içerik vektörleri oluşturulmakta ve içerik vektörleri gelecekteki yamaların temsillerinin tahmininde kullanılmaktadır. C içerik vektörü ve Z_{t+k} tahmin edilecek yamanın temsili olmak üzere CPC [52] C ve Z_{t+k} arasındaki karşılıklı bilgiyi ($I(C; Z_{t+k})$) maksimize etmeye çalışarak görüntü temsilini öğrenmeye çalışmaktadır. Ancak yüksek boyutlu rastgele değişkenler arasındaki karşılıklı bilginin hesaplanma zorluğu nedeniyle, karşılıklı bilginin alt sınırının tahmininde InfoNCE [52] kullanılmaktadır. Bu kayıp fonksiyonunda içerik vektörü kullanılarak elde edilen tahmini yama temsili ve gelecekteki yamanın temsili pozitif çifti oluştururken, negatif çift diğer görüntülerin yama temsilleri ile oluşturulmaktadır. InfoNCE kayıp fonksiyonu, kodlayıcı ve otoregresif modeli gelecekteki yama temsillerine benzer tahmin yapmaya zorlamaktadır. CPC [52] mimarisinde yapılan değişiklikle sadece görüntü verisine uygulanan az etiketli veride başarılı olan CPC-v2 [55] geliştirilmiştir. CPC-v2 [55] mimarisinde daha derin ve geniş kodlayıcı kullanılmaktadır. Şekil 13'te CPC [52] yöntemi gösterilmiştir.



Şekil 13. *CPC [52] yönteminin gösterimi*

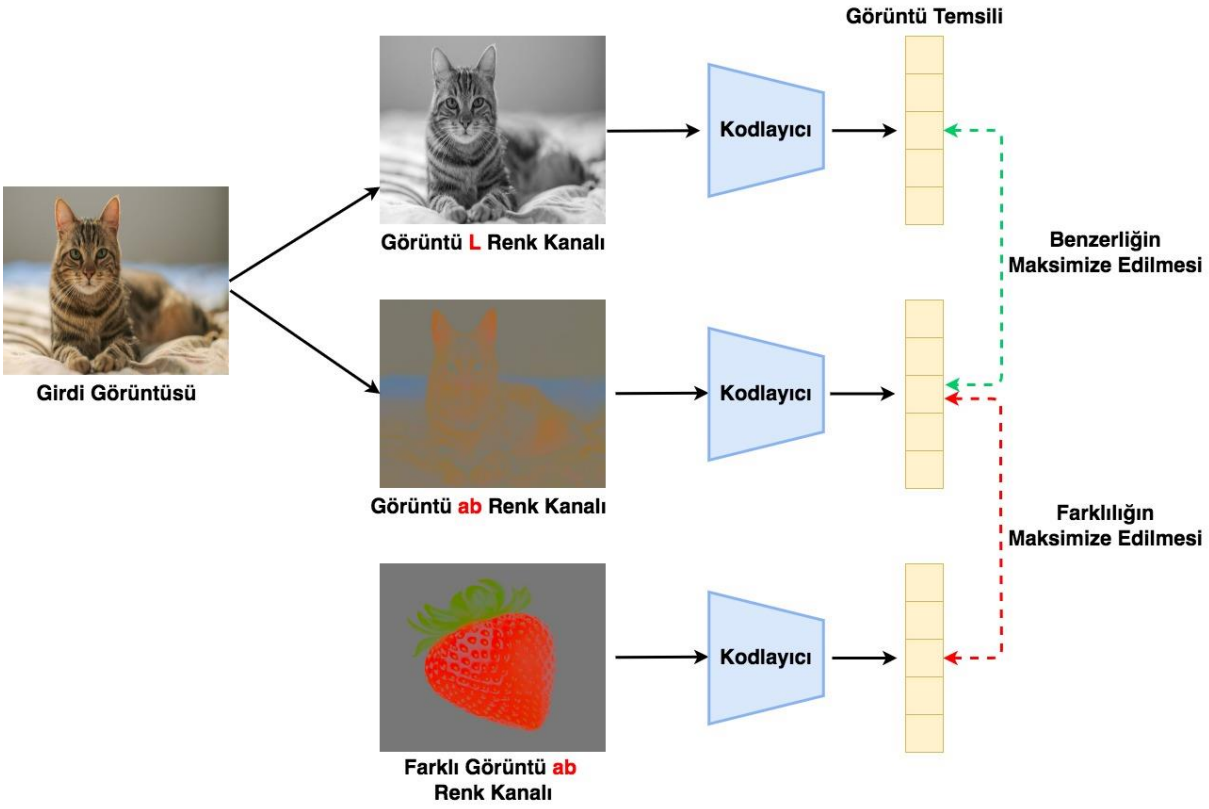
DIM [56] yönteminde CPC [52] yöntemine benzer şekilde InfoNCE [52] kayıp fonksiyonu kullanılarak global ve yerel öznitelikler arasındaki karşılıklı bilgi maksimize edilmeye çalışılmaktadır. Bu çalışmada görüntü temsili global ve yerel öznitelik çiftinin aynı görüntüden olup olmadığını sınıflandırarak öğrenilmektedir. Bir görüntünün kodlayıcıdan geçirilerek elde edilen son çıktı global öznitelikleri oluştururken, yerel öznitelikler ise kodlayıcının ara katmanların çıktılarından oluşmaktadır. Karşılaştırmalı kayıp fonksiyonunda, aynı görüntüye ait global ve yerel öznitelikler pozitif çifti oluştururken, farklı görüntülere ait global ve yerel öznitelikler negatif çifti oluşturmaktadır. Bu kayıp fonksiyonunun minimize edilebilmesi için global öznitelik vektörünün tüm farklı yerel bölgelerden bilgiyi yakalaması gerekmektedir. DIM [56] yönteminden farklı olarak, daha güçlü kodlayıcı ve girdi görüntüsünün çoğaltılmış görünümlerinden oluşturulan pozitif çiftleri kullanan AMDIM [57] geliştirilmiştir. Bu yöntemde, bir görüntünün çoğaltılmış görünümlerinin çoklu ölçeklerinden çıkarılan yerel öznitelikler ve çoğaltılmış görünümlerinin global öznitelikleri arasındaki

karşılıklı bilgiyi maksimize etmeye çalışarak görüntü temsili öğrenilmektedir. Şekil 14'te AMDIM [57] yöntemi gösterilmiştir.



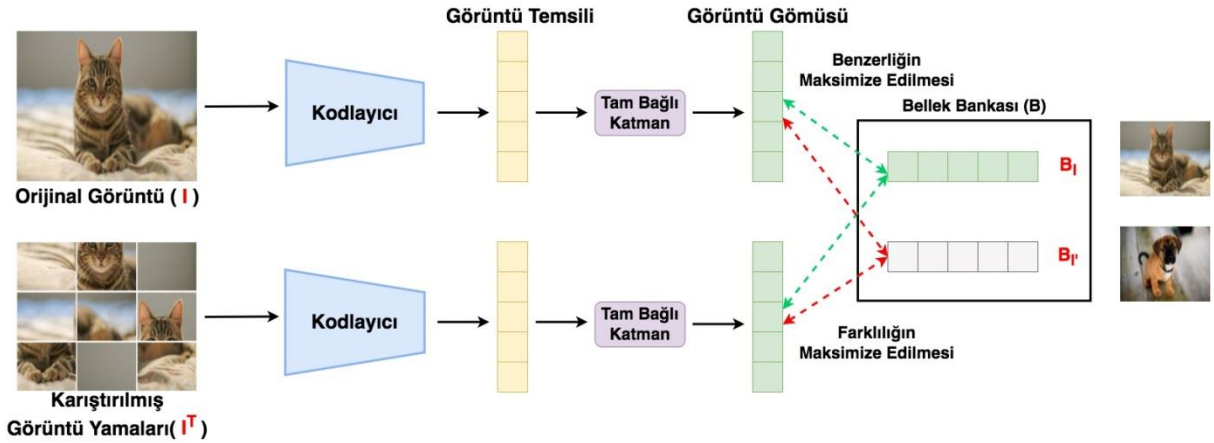
Şekil 14. *AMDIM [57] yönteminin gösterimi*

Diğer karşılaştırmalı öz-denetimli öğrenme yöntemlerinden farklı bir şekilde pozitif ve negatif çiftlerin oluşturulduğu CMC [58] yönteminde, bir görüntünün farklı görünümlerinin temsilleri arasında karşılıklı bilgiyi maksimize etmeye çalışarak genelleştirilebilir görüntü temsili elde edilmeye çalışılmaktadır. Pozitif çift, aynı görüntünün farklı görünümleriyle oluşturulurken, negatif çift ise farklı görüntülerin görünümleriyle oluşturulmaktadır. Örneğin *Lab* renk uzayındaki bir görüntünün *L* kanalı, aynı görüntünün *ab* kanalı ile pozitif çifti oluştururken başka bir görüntünün *ab* kanalı ile negatif çifti oluşturmaktadır. Ayrıca bu çalışmada ikiden fazla görünüm kullanılarak öğrenilen görüntü temsilinin kalitesinin arttığı gösterilmiştir. Şekil 15'te CMC [58] yöntemi gösterilmiştir.



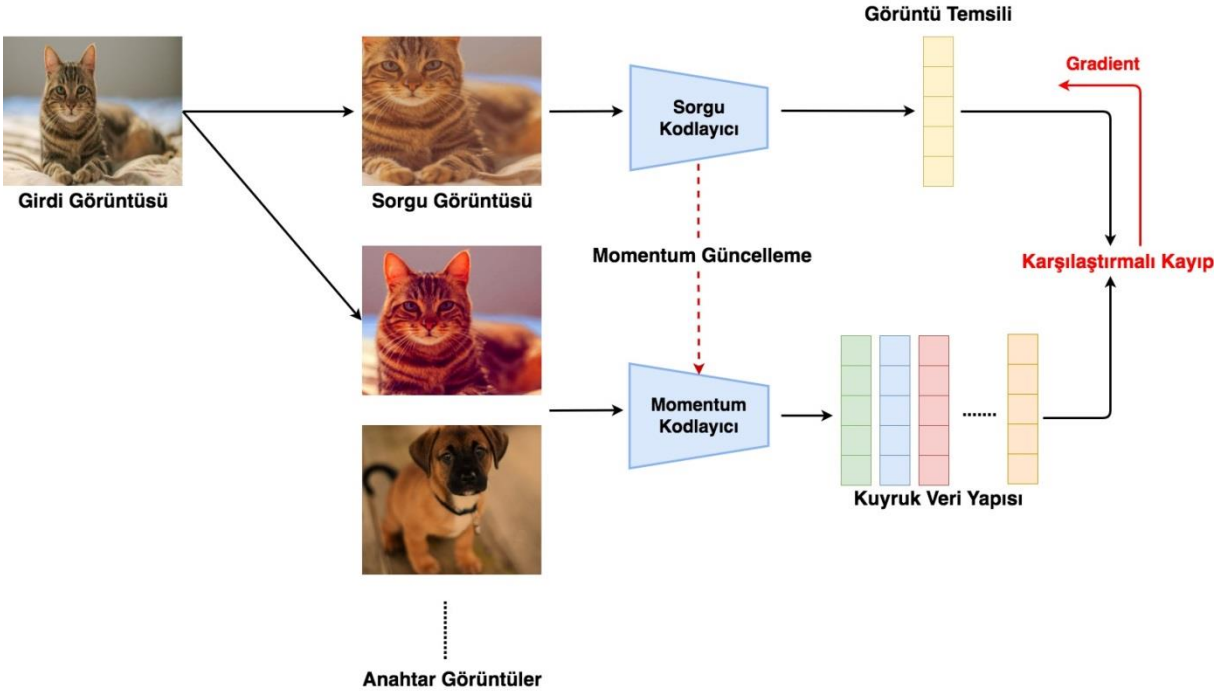
Şekil 15. *CMC [58] yönteminin gösterimi*

Karşılaştırmalı öğrenme ve yardımcı görevin birlikte kullanıldığı PIRL [59] yöntemi, yardımcı görevden bağımsız genelleştirilebilir görüntü temsilleri elde edebilmektedir. Yardımcı görevler kullanarak görüntü temsilini öğrenmeye çalışılan öz-denetimli öğrenme yöntemleri, yardımcı göreve özgü görüntü temsillerini öğrendikleri için bu görüntü temsillerinin hedef görevler için genelleştirilmesinde problemler yaşanmaktadır. PIRL [59] yönteminde pozitif çift, orijinal görüntü ve görüntünün döndürülmesiyle veya yapboz bulmacaya dönüştürülmesiyle elde edilen görüntüyle oluşturulurken, negatif çift ise orijinal görüntü ve veri setindeki farklı görüntüler kullanılarak oluşturulmaktadır. Görüntüye uygulanan bu dönüşüm görüntüde anlamsal değişikliğe neden olmayacağı için pozitif çiftlerden elde edilen temsiller temsil uzayında birbirine yakın olmaya zorlanırken negatif çiftler ise birbirinden uzak olmaya zorlanmaktadır. Karşılaştırmalı öğrenmede genelleştirilebilir temsilin elde edilebilmesi için çok sayıda negatif örnek gerektiğinden, bu çalışma negatif örneklerin temsillerini InstDisc [50] çalışmasına benzer şekilde çok sayıda negatif çift tutabilen bellek bankasında tutmaktadır. Her mini-yığında hesaplanan görüntü temsillerinin bellek bankasındaki karşılıkları üstel hareketli ortalamayla güncellenmektedir. Şekil 16'da PIRL [59] yöntemi gösterilmiştir.



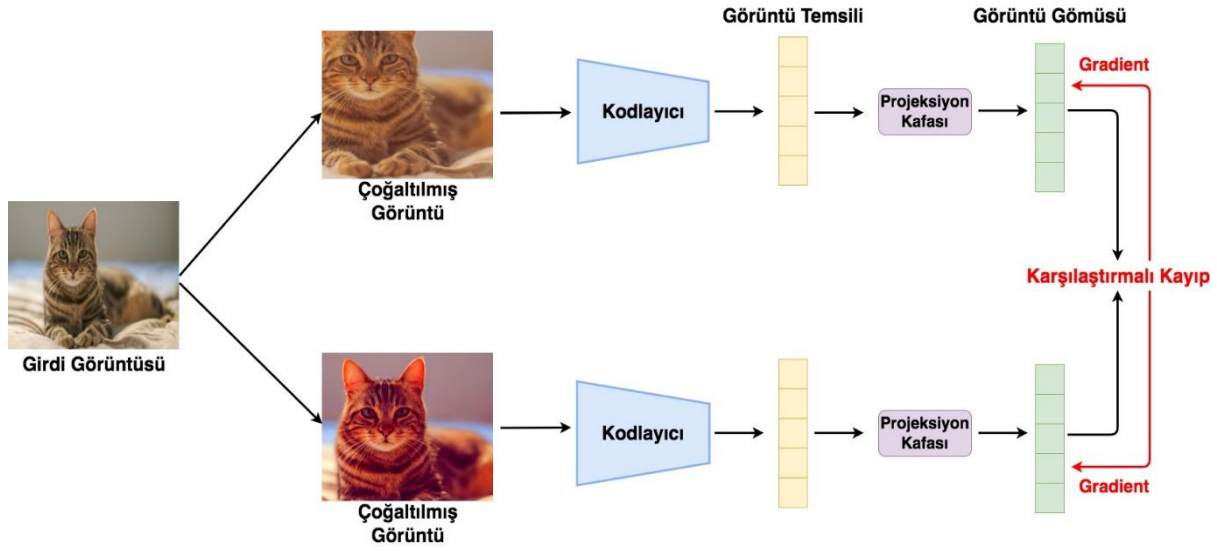
Şekil 16. *PIRL [59] yönteminin gösterimi*

İki kodlayıcıya sahip karşılaştırmalı öğrenme yöntemi olan MoCO [26] yönteminde negatif örnekler için kuyruk veri yapısı kullanılmaktadır. Bu yöntemde, her girdi görüntüsüne iki farklı veri çoğaltma uygulanarak sorgu ve anahtar olarak adlandırılan iki dönüştürülmüş görüntü oluşturulmaktadır. Sorgu ve anahtar görüntüleri, sorgu kodlayıcı ve momentum kodlayıcı olarak adlandırılan iki farklı kodlayıcıdan geçirilerek görüntü temsilleri elde edilmektedir. MoCO [26] yönteminde karşılaştırmalı kayıp fonksiyonun çok sayıda negatif örnek gereksinimi için negatif örnekler PIRL [59] ve InstDisc [50] yöntemlerinde kullanılan bellek bankasından farklı olarak kuyruk veri yapısında tutulmaktadır. Kuyruk veri yapısında görüntü temsillerinin tutarlı olmasını sağlamak için anahtar görüntülerinin geçerli mini-yığındaki görüntü temsilleri kuyruğa eklenirken, çok önceden kuyruğa eklenmiş güncelliğini yitirmiş görüntü temsilleri kuyruktan çıkarılmaktadır. Momentum kodlayıcının ağırlıkları, sorgu kodlayıcının ağırlıklarının üstel hareketli ortalamasıyla yavaş bir şekilde güncellenerek görüntü temsillerinin tutarlı olması sağlanmaktadır. MoCO [26] yöntemine çok katmanlı algılayıcı kafası ve daha güçlü veri çoğaltma yöntemleri eklenerek MoCO-v2 [60] geliştirilmiştir. Şekil 17'de MoCO [26] yöntemi gösterilmiştir.



Şekil 17. *MoCO [26] yönteminin gösterimi*

SimCLR [25], güçlü veri çoğaltma tekniklerinin ve negatif örnekler için yüksek boyutlu yığının kullanıldığı diğer bir karşılaştırmalı öğrenme yöntemidir. Negatif örnekler, PIRL [59] ve InstDisc [50] yöntemlerinde bellek bankasında, MoCO [26] yönteminde kuyruk veri yapısında tutulmaktayken, SimCLR [25] yönteminde ise negatif örnekler yüksek hesaplama maliyetine rağmen büyük boyutlu yığında tutulmaktadır. Bu yöntemde ilk olarak yığındaki her bir görüntüye rastgele seçilen iki farklı veri çoğaltma tekniği uygulanarak pozitif çiftler oluşturulmaktadır. Bu dönüştürülmüş görüntüler kodlayıcı ve doğrusal olmayan projeksiyon kafasından geçirilerek gömü vektörleri elde edilmektedir. Karşılaştırmalı kayıp fonksiyonu, pozitif çiftlerden elde edilen gömü vektörlerini gömü uzayında birbirine yakın olmaya zorlarken yığındaki diğer dönüştürülmüş görüntülerin gömü vektörlerinden ise uzak olmaya zorlamaktadır. SimCLR [25] mimarisinde değişiklikler yapılarak SimCLR-v2 [61] geliştirilmiştir. SimCLR-v2 [61] mimarisinde daha büyük kodlayıcı, daha büyük yığın ve daha derin doğrusal olmayan projeksiyon kafası kullanılmaktadır. Şekil 18'de SimCLR [25] yöntemi gösterilmiştir.



Şekil 18. *SimCLR [25] yönteminin gösterimi*

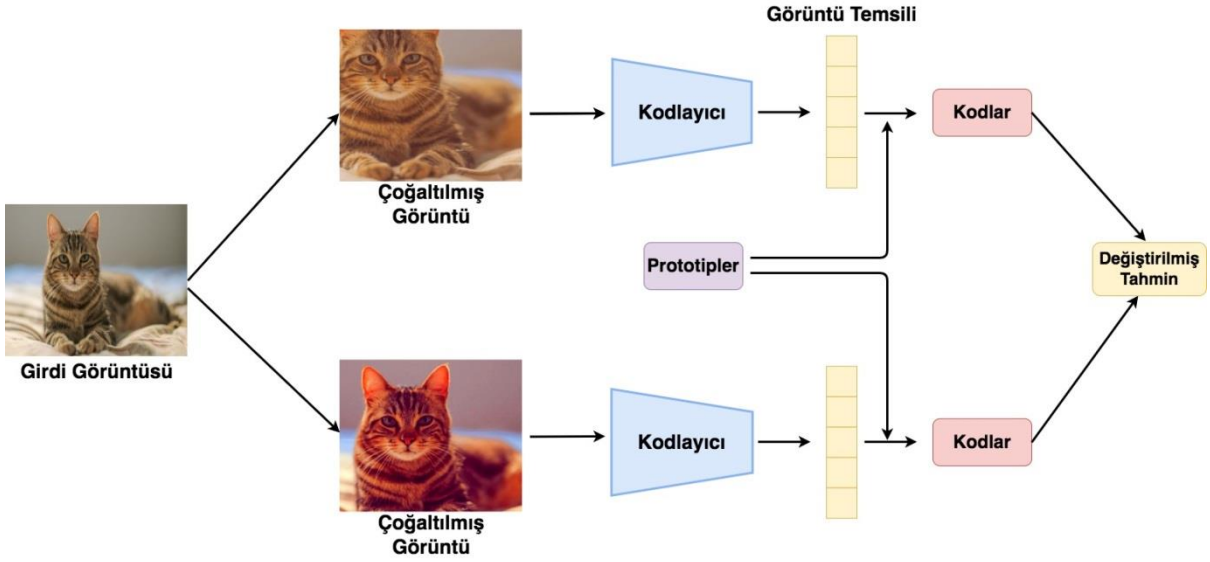
SwAV [30], çevrim içi kümelemeyi ve karşılaştırmalı öğrenmeyi birlikte kullanarak görüntü temsillerini öğrenmeye çalışan bir yöntemdir. Çevrim dışı kümelemede sahte etiketlerin elde edilmesi için tüm veri seti üzerinden en az bir defa geçilmesi gerekliyken, çevrim içi kümelemede mini-yığındaki görüntüler kümelenmekte ve DESA'nın eğitimi için gerekli olan sözde etiketler elde edilmektedir. Böylece kümeleme ve öğrenme faaliyetleri için tek bir kayıp fonksiyonu kullanılmaktadır. Bu yöntemin amacı, diğer karşılaştırmalı öğrenme yöntemlerindeki gibi sadece pozitif çiftleri birbirine yakınlaştırmak değil, aynı zamanda birbirine benzeyen diğer tüm görüntü temsillerinin bir araya gelmesini sağlamaktır. Bu çalışmada veri setini özetleyen ve her biri bir kümeye karşılık gelen eğitilebilir prototip vektörleri kullanılmaktadır. Diğer karşılaştırmalı öğrenme yöntemlerine benzer şekilde girdi olarak görüntülerin veri çoğaltmayla oluşturulmuş görünümleri alınmaktadır. Oluşturulan görünümlerden elde edilen temsillerin prototip vektörlerine atanması ile kümeleme gerçekleştirilmektedir. Kümeleme işlemi en uygun aktarım problemlerinin çözümünde kullanılan Sinkhorn-Knopp algoritmasıyla [45] gerçekleştirilmektedir. Sinkhorn-Knopp algoritması [45], temsil vektörlerinin oluşturduğu matris ile prototip vektörlerinin oluşturduğu matrisin çarpımıyla elde edilen maliyet matrisi ve her kümeye eşit sayıda görüntü temsili atanması kısıtını kullanarak kümelemeyi gerçekleştirmektedir. Görüntü temsillerinin kümelere atanmasıyla DESA'nın eğitimi için gerekli sözde etiketler elde edilmiştir. Bu yöntem, pozitif çiftlerin aynı kümeye atanmasını zorlamak için kayıp fonksiyonunda bu iki görünüm temsilinin sözde etiketlerini yer değiştirmektedir. Bir görüntünün görünümlerinin temsilleri z_s , z_t ve temsillerin sözde etiketleri q_s , q_t olmak üzere, görüntü temsillerinin kayıp fonksiyonu $\ell(z_t, q_s)$, $\ell(z_s, q_t)$ ve toplam kayıp $L(z_t, z_s)$ olarak gösterilmektedir. Eşt. 4'te de görüleceği üzere bir görünümün temsilinin sözde etiketi diğer görünümün temsilleri kullanılarak tahmin edilmeye çalışılmaktadır.

$$L(z_t, z_s) = \ell(z_t, q_s) + \ell(z_s, q_t)$$
$$\ell(z_t, q_s) = -\sum_k q_s^{(k)} log p_t^{(k)} \tag{4}$$

Diğer karşılaştırmalı öğrenme yöntemlerinde doğrudan görüntü temsilleri karşılaştırırken, Eşt. 5'te görüleceği üzere bu yöntemde görüntü temsilleri prototip vektörleri ile karşılaştırılmaktadır. k küme sayısını, τ sıcaklık derecesini ve $p_t^{(k)}$ise görüntü temsilinin k. küme prototipine benzeme olasılığını göstermektedir.

$$p_t^{(k)} = \frac{\exp(\frac{1}{\tau} z_t^T c_k)}{\sum_{k'} \exp(\frac{1}{\tau} z_t^T c_{k'})} \tag{5}$$

Şekil 19'da SwAV [30] yöntemi gösterilmiştir.



Şekil 19. *SwAV [30] yönteminin gösterimi*

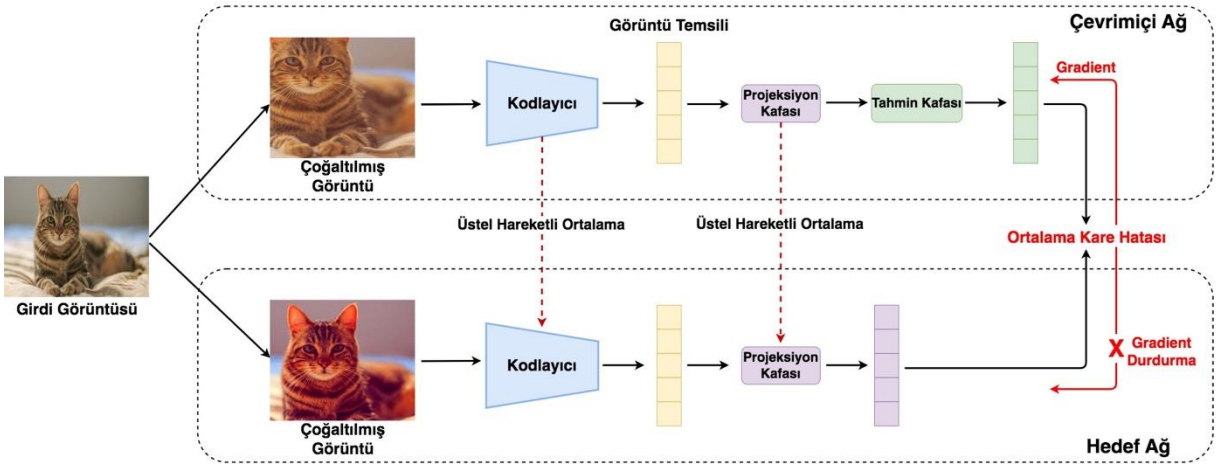
E. KARŞILAŞTIRMALI OLMAYAN ÖZ-DENETİMLİ ÖĞRENME

Aynı ağırlıkların paylaşıldığı siyam ağlarının eğitiminde, sadece pozitif çiftlerin kullanımı bu ağların tüm girdi görüntüleri için sabit bir vektör üretmesine neden olmaktadır. Genelleştirilebilir görüntü temsillerinin elde edilemediği bu durum başarısız çözüm olarak adlandırılmaktadır. Bu nedenle bu sorunun çözümü için karşılaştırmalı öğrenme yöntemlerinin eğitiminde çok sayıda negatif çift kullanılmaktadır. Karşılaştırmalı olmayan öz-denetimli öğrenme yöntemleri, negatif çiftleri kullanmadan genelleştirilebilir görüntü temsillerini elde edebilen yöntemlerdir.

BYOL [27] yöntemi, negatif çift kullanmadan görüntü temsilinin öğrenilmesini gerçekleştirebilen ilk karşılaştırmalı olmayan öz-denetimli öğrenme yöntemidir. Bu yöntem birbiriyle etkileşime giren ve birbirinden öğrenen çevrim içi ve hedef olarak adlandırılan iki ağı içermektedir. Çevrim içi ağ, kodlayıcı, projeksiyon kafası ve tahmin kafasını içerirken, hedef ağ ise kodlayıcı ve projeksiyon kafasından oluşmaktadır. Hedef ağın ağırlıkları çevrim içi ağın üstel hareketli ortalamasıyla güncellenmektedir. Bu yöntemde, bir görüntüye iki farklı veri çoğaltma tekniği uygulanarak elde edilen dönüştürülmüş görüntüler çevrim içi ağa ve hedef ağa girdi olarak verilmekte ve çevrim içi ağ, hedef ağdan elde edilecek temsili tahmin edecek şekilde eğitilmektedir. Bu iki ağ tarafından elde edilen temsillerin temsil uzayındaki mesafesini minimize etmek için karşılaştırmalı kayıp fonksiyonundan farklı olarak, ortalama kare hatası kullanılmaktadır. Çevrim içi ağ ve hedef ağdan elde edilen gömü vektörleri sırasıyla $q_\theta(z_\theta)$ ve z'_ξ olmak üzere, ilk olarak gömü vektörlerine L2 normalizasyon uygulanmakta, sonrasında ise normalize edilen gömü vektörlerinin skaler çarpımıyla Eşt. 6'da verildiği şekilde BYOL [27] kayıp fonksiyonu hesaplanmaktadır.

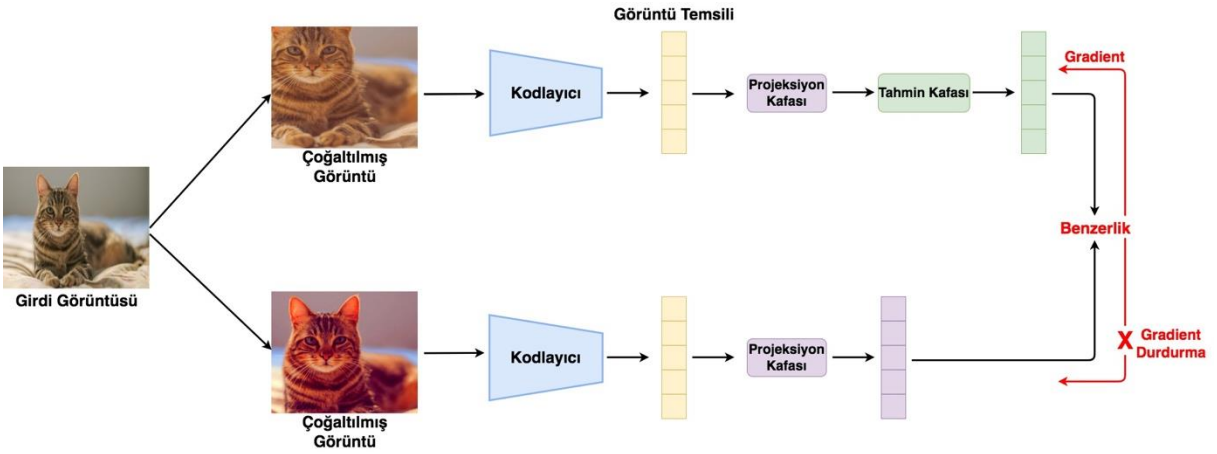
$$\mathcal{L}_\theta^{BYOL} = 2 - 2.\frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 . \left\|z'_\xi\right\|_2} \tag{6}$$

Bu yöntemin diğer karşılaştırmalı öğrenme yöntemlerine kıyasla yığın boyutundaki ve görüntü çoğaltma tekniklikleriindeki değişikliklere karşı daha dayanıklı olduğu gösterilmiştir. Şekil 20'de BYOL [27] yöntemi gösterilmiştir.



Şekil 20. *BYOL [27] yönteminin gösterimi*

Diğer bir karşılaştırmalı olmayan öz-denetimli öğrenme yöntemi SimSiam [28], BYOL [27] yöntemine benzer şekilde negatif çiftleri kullanmadan görüntü temsilinin öğrenildiği bir yöntemdir. Başarısız çözüm problemini önlemek için ağın bir tarafında tahmin kafası ve ağın diğer tarafında ise gradientin durdurulması işleminin kullanıldığı SimSiam [28] yöntemi, momentum kodlayıcı kullanmayan BYOL [27] yöntemi olarak düşünülebilir. Bu yöntemde negatif çiftler ve momentum kodlayıcı kullanılmadan, aynı görüntünün çoğaltılmış iki görüntüsü girdi olarak alınmakta ve elde edilen görüntü temsillerinin benzerliği maksimize edilmeye çalışılmaktadır. Şekil 21'de SimSiam [28] yöntemi gösterilmiştir.



Şekil 21. *SimSiam [28] yönteminin gösterimi*

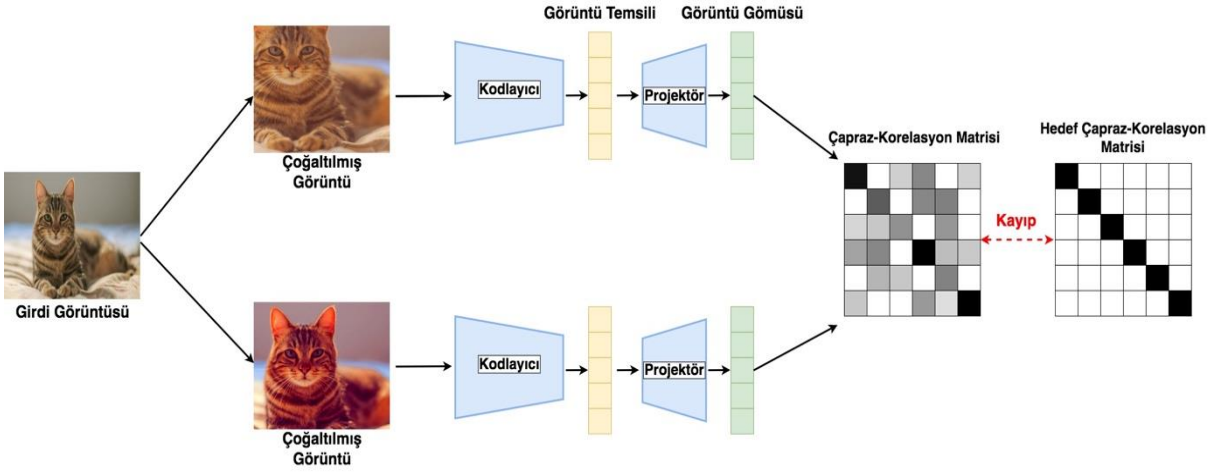
Görüntü temsili öğreniminde negatif çiftlerin kullanılmadığı diğer bir yöntem Barlow Twins [29] yöntemidir. Ancak BYOL [27] ve SimSiam [28] yöntemleri, başarısız çözümü önlemek için tahmin edici kafasına ve gradientin durdurulması işlemine ihtiyaç duyarken, Barlow Twins [29] yöntemi farklı bir kayıp fonksiyonu kullanarak başarısız çözümü önleyebilmektedir. Bu yöntemde, yığındaki (batch) her bir görüntünün veri çoğaltmayla elde edilen iki görünümü aynı ağırlıklara sahip ikiz ağlardan geçirilerek z^A ve z^B görüntü temsilleri matrisi (yığın boyutu x görüntü temsili boyutu) oluşturulmaktadır. Bu matrislerin sütunlarının skaler çarpımıyla C kare çapraz-korelasyon matrisinin her bir değeri hesaplanmaktadır. b yığını, i ve j görüntü temsillerinin boyut indislerini temsil etmek üzere Eşt. 7'de C çapraz-korelasyon matrisi verilmiştir.

$$C_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}} \tag{7}$$

Oluşturulan çapraz-korelasyon matrisini birim matrise mümkün olduğunca yakın hale getirmeyi amaçlayan bir kayıp fonksiyonu kullanılmaktadır. Bu kayıp fonksiyonu sayesinde bir görüntünün veri çoğaltmayla elde edilen iki görünümünden elde edilen temsiller birbirine benzer hale getirilirken, temsil vektörlerinin boyutları arasındaki korelasyon minimize edilmektedir. İki terimden oluşan kayıp fonksiyonunda λ, kayıp fonksiyonun birinci ve ikinci terimlerinin önemini değiştiren bir parametredir. Eşt. 8'de, Barlow Twins [29] yönteminde kullanılan kayıp fonksiyonu verilmiştir.

$$\mathcal{L}_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} {C_{ij}}^2 \tag{8}$$

Şekil 22'de Barlow Twins [29] yöntemi gösterilmiştir.



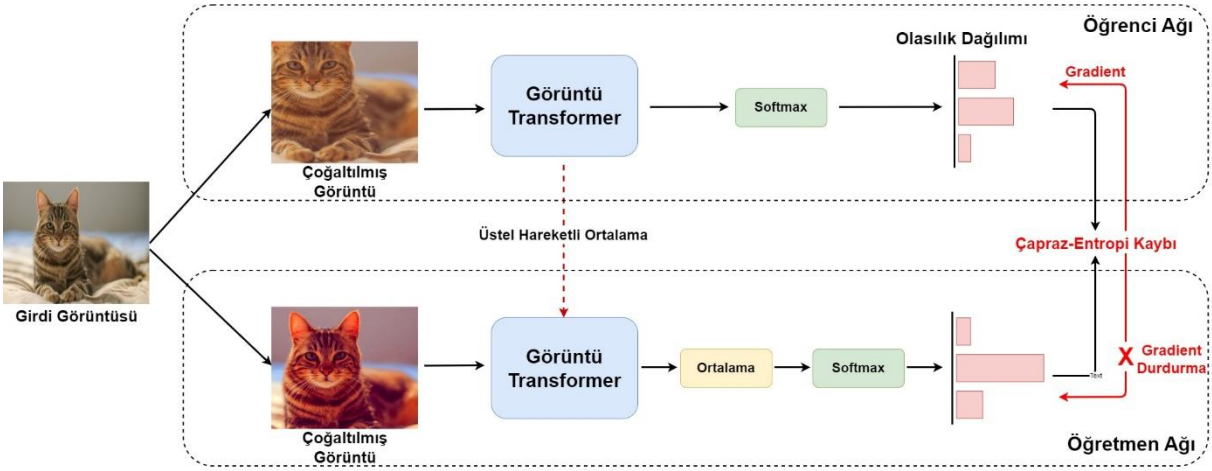
Şekil 22. *Barlow Twins [29] yönteminin gösterimi*

SimSiam [28], BYOL [27] yöntemlerine benzer şekilde, birbiriyle etkileşime giren ve birbirinden öğrenen öğrenci ve öğretmen olarak adlandırılan iki ağa sahip olan DINO [62] , negatif çiftleri kullanmadan görüntü temsilinin öğrenildiği diğer bir yöntemdir. Diğer öz-denetimli öğrenme yöntemlerinden farklı olarak esnek bir yapıya sahip olan DINO [62] yönteminde, ön-eğitimli ağ mimarisi olarak derin evrişimsel sinir ağları veya görüntü transformer mimarisi [63] kullanılabilmektedir. Aynı derin evrişimsel sinir ağ veya görüntü transformer mimarine sahip olan öğrenci ve öğretmen ağlarının ağırlıkları birbirinden farklıdır. Öğretmen ağının ağırlıkları öğrenci ağının üstel hareketli ortalamasıyla güncellenmektedir. DINO [62] yönteminde çoklu ölçekli kırpma veri çoğaltma tekniği uygulanmaktadır. Öğretmen ağı girdi olarak görüntünün büyük bir kısmını içeren global çoğaltılmış görüntüleri alırken, öğrenci ağı ise girdi olarak global çoğaltılmış görüntülerle birlikte görüntünün küçük bir kısmını içeren lokal görüntüleri de almaktadır. Çoğaltılmış görüntülerin öğrenci ağı çıktısına doğrudan softmax fonksiyonu uygulanırken, öğretmen ağı çıktısına softmax fonksiyonundan önce ortalama (centering) işlemi uygulanmaktadır. Ayrıca öğretmen ağına uygulanan softmax fonksiyonunda kullanılan sıcaklık derecesi, öğrenci ağına kıyasla daha küçük seçilerek keskinleştirme (sharpening) yapılmaktadır. Öğretmen ağına uygulanan ortalama ve keskinleştirme işlemleriyle başarısız çözüm problemini önlemek amaçlanmaktadır. Bu yöntemde öğrenci ve öğretmen ağını benzer tahminler yapmaya zorlayan çapraz-entropi kullanılmaktadır. DINO [62] yöntemine benzer şekilde görüntü transformer mimarisini [63] kullanan ve ince ayar yapma ihtiyaç duymadan görüntü sınıflandırma, görüntü bölütleme, nesne tespiti, derinlik tahmini vb. bilgisayarlı görü uygulamalarında yüksek başarı elde eden DINOv2 [64] yöntemi geliştirilmiştir. DINO [62] yönteminin ön-eğitiminde sadece ImageNet [20] veri seti kullanılırken, DINOv2 [64]

yönteminin ön-eğitimi için çeşitli veri setleri ve web kaynakları kullanılarak 142 milyon görüntüden oluşan çok daha büyük bir veri seti oluşturulmuştur (LVD-142M) ve uygulamada çeşitli iyileştirmeler yapılmıştır. Öğrenci ağının parametreleri θ_s, öğretmen ve öğrenci ağlarının olasılık dağılımları sırasıyla P_t ve P_s ve $H(a,b) = -a.logb$ olmak üzere DINO [62] kayıp fonksiyonu Eşt. 9'da verilmiştir.

$$min_{\theta_s} H\big(P_t(x), P_s(x)\big) \tag{9}$$

Şekil 23'te DINO [62] yöntemi gösterilmiştir.



Şekil 23 *DINO [62] yönteminin gösterimi*

III. PERFORMANS DEĞERLENDİRMESİ

Öz-denetimli öğrenme yöntemleriyle öğrenilen görüntü temsillerinin kalitesinin değerlendirilmesinde genellikle doğrusal değerlendirme ve ince-ayar kullanılmaktadır [39], [65].

A. DOĞRUSAL DEĞERLENDİRME

Doğrusal değerlendirmede öz-denetimli öğrenme yöntemleriyle öğrenilen ön-eğitimli ağdan sonra bir doğrusal sınıflandırıcı kullanılmaktadır. Oluşturulan bu mimaride etiketli veri ile doğrusal sınıflandırıcı eğitilirken ön-eğitimli ağın ağırlıkları dondurulmaktadır. Böylece öz-denetimli öğrenme yöntemleriyle öğrenilen görüntü temsillerinin doğrusal olarak ne kadar ayrılabilir olduğu ölçülebilmektedir. İlk öz-denetimli öğrenme yöntemlerinde ön-eğitimli ağ mimarisi olarak AlexNet [3] omurgası kullanılmaktayken son yıllarda önerilen öz-denetimli öğrenme yöntemlerinde ön-eğitimli ağ mimarisi olarak ResNet [5] ve görüntü tansformer omurgaları [63] kullanılmaktadır. Bu nedenle öz-denetimli yöntemlerin doğrusal değerlendirmesinde, ön-eğitimli ağ mimarisi olarak AlexNet [3] omurgasının kullanıldığı öz-denetimli yöntemler, ResNet [5] ve görüntü tansformer omurgalarının [63] kullanıldığı öz-denetimli yöntemler üç tabloda verilmiştir. Tablo 1 ve Tablo 2'de AlexNet [3] ve ResNet [5] mimarileri kullanılarak etiketsiz ImageNet [20] veri setinde ön-eğitilen öz-denetimli öğrenme yöntemlerinin, ImageNet [20] ve Places205 [66] veri setlerinde doğrusal sınıflandırma Top-1 doğruluğu verilmektedir. Tablo 3'te ise ImageNet [20] veri setinde ön-eğitilen DINO [62] ve LVD-142M veri setinde ön eğitilen DINO-v2 yöntemlerinin ImageNet [20] veri setindeki doğrusal sınıflandırma Top-1 doğruluğu verilmektedir. Görüldüğü üzere yeniden oluşturma temelli, tahmin temelli ve kümeleme temelli öz-denetimli öğrenme yöntemleri, ön-eğitimli ağ mimarisi olarak ResNet [5] omurgası kullanıldığında daha yüksek performans göstermelerine rağmen, karşılaştırmalı ve karşılaştırmalı olmayan öz-denetimli öğrenme yöntemleriyle rekabet edecek sonuçlar elde

edememişlerdir. Ayrıca Tablo 3'te görüldüğü üzere öz denetimli öğrenme yöntemlerinde görüntü transformer mimarileri kullanımı gelecek vadeden bir konudur.

Tablo 1. *AlexNet [3] mimarisi kullanılarak, etiketsiz ImageNet [20] veri setinde ön-eğitilen öz-denetimli öğrenme yöntemlerinin, ImageNet [20] ve Places205 [66] veri setlerinde doğrusal sınıflandırma Top-1 doğruluğu. [39]–[41] yöntemlerin Top-1 doğruluğu değerleri [35] yayınından alınmıştır.*

Yöntem	Kategori	ImageNet	Places205
Denetimli Öğrenme [3]	-	50.5	39.4
Renklendirme [39]	Yeniden oluşturma temelli	32.6	30.3
Görece Pozisyon Tahmini [40]	Tahmin temelli	31.7	32.7
Yapboz Bulmaca [41]	Tahmin temelli	34.7	35.5
Döndürme [42]	Tahmin temelli	38.7	35.1
Derin Kümeleme [43]	Kümeleme temelli	39.8	37.5

Tablo 2. *ResNet [5] mimarileri kullanılarak, etiketsiz ImageNet [20] veri setinde ön-eğitilen öz-denetimli öğrenme yöntemlerinin, ImageNet [20] ve Places205 [66] veri setlerinde doğrusal sınıflandırma Top-1 doğruluğu. [39], [41] yöntemlerin Top-1 doğruluğu değerleri [65] yayınından alınmıştır.*

Yöntem	Kategori	ImageNet	Places205
Denetimli Öğrenme [25]	-	76.5	53.2
Renklendirme [39]	Yeniden oluşturma temelli	39.6	37.5
Yapboz Bulmaca [41]	Tahmin temelli	45.7	41.2
Döndürme [42]	Tahmin temelli	48.9	41.4
Derin Kümeleme [43]	Kümeleme temelli	-	45.5
SeLA [44]	Kümeleme temelli	61.5	-
IntDisc [50]	Karşılaştırmalı	54	-
CPC [52]	Karşılaştırmalı	48.7	-
CPC-v2 [55]	Karşılaştırmalı	63.8	-
PIRL [59]	Karşılaştırmalı	63.6	49.8
MoCO [26]	Karşılaştırmalı	60.6	48.9
MoCO-v2 [60]	Karşılaştırmalı	71.1	51.8
SimCLR [25]	Karşılaştırmalı	69.3	52.5
SwAV [30]	Karşılaştırmalı	75.3	56.7
BYOL [27]	Karşılaştırmalı olmayan	74.3	54
SimSiam [28]	Karşılaştırmalı olmayan	71.3	-
Barlow Twins [29]	Karşılaştırmalı olmayan	73.2	54.1
DINO [62]	Karşılaştırmalı olmayan	75.3	-

Tablo 3. *GT-T/8, 8 x 8 yama büyüklüğü temel görüntü transformer ve GT-B/14, 14 x 14 yama büyük görüntü transformer olmak üzere, ImageNet [20] veri setinde ön-eğitilen DINO [62] ve LVD-142M veri setinde ön eğitilen DINO-v2 yöntemlerinin ImageNet [20] veri setindeki doğrusal sınıflandırma Top-1 doğruluğu.*

Yöntem	Kategori	ImageNet
Denetimli Öğrenme [63]	-	79.9
DINO [62] (GT-T/8)	Karşılaştırmalı olmayan	80.1
DINO-v2 [64] (GT-B/14)	Karşılaştırmalı olmayan	86.5

B. İNCE AYAR

Öz-denetimli öğrenme yöntemleriyle öğrenilen görüntü temsillerinin kalitesinin değerlendirilmesinde kullanılan diğer yöntem ince ayardır. Bu yöntemde ön-eğitimli ağın ağırlıkları hedef görevde başlangıç ağırlıkları olarak kullanılmakta ve doğrusal değerlendirmeden farklı olarak eğitim sırasında bu ağırlıklara ince ayar yapılmaktadır. Ayrıca doğrusal değerlendirmede performans değerlendirmesi

sadece görüntü sınıflandırma görevi için yapılabilmekteyken, bu yöntem sayesinde nesne tespiti, görüntü bölütleme vb. bilgisayarlı görü görevleri performans değerlendirmesinde kullanılabilmektedir. Böylece öz-denetimli öğrenmeyle elde edilen görüntü temsillerinin farklı görevlere genelleştirilebilirliği değerlendirilebilmektedir. Doğrusal değerlendirmede olduğu gibi bu yöntemde de ön-eğitimli ağ mimarisi olarak AlexNet [3] omurgasının kullanıldığı öz-denetimli yöntemler ve ResNet [5] omurgasının kullanıldığı öz-denetimli yöntemler iki tabloda verilmiştir. Tablo 4'te AlexNet [3] mimarisi kullanılarak etiketsiz ImageNet [20] veri setinde ön eğitilen öz-denetimli öğrenme yöntemlerinin, Pascal VOC 2007 [67] veri setinde görüntü sınıflandırma ve nesne tespiti hedef görevleri, Pascal VOC 2012 [68] veri setinde ise görüntü bölütleme hedef görevi için performans değerlendirmesi verilmiştir. Tablo 5'te ResNet [5] mimarisi kullanılarak etiketsiz ImageNet [20] veri setinde ön eğitilen öz-denetimli öğrenme yöntemlerinin, Pascal VOC 2007 [67] veri setinde nesne tespiti hedef görevi için performans değerlendirmesi verilmiştir. Görüldüğü üzere karşılaştırmalı ve karşılaştırmalı olmayan öz denetimli öğrenme yöntemleri denetimli öğrenmeyle rekabet edecek sonuçlar elde etmişlerdir.

Tablo 4. *AlexNet [3] mimarisi kullanılarak, Pascal VOC 2007 [67] veri setinde görüntü sınıflandırma, nesne tespiti hedef görevleri ve Pascal VOC 2012 [68] veri setinde görüntü bölütleme hedef görevi performans değerlendirmesi. Görüntü sınıflandırma ve nesne tespitinde ortalama AP ve görüntü bölütlemede ise ortalama IOU performans değerlendirme metriği kullanılmıştır. Nesne tespiti ve görüntü bölütleme hedef görevlerinde, omurga olarak AlexNet [3] kullanan sırasıyla Fast R-CNN [11] ve FCN [12] mimarileri kullanılmıştır.*

Yöntem	Kategori	Sınıflandırma	Nesne Tespiti	Görüntü Bölütleme
Denetimli Öğrenme [3]	-	79.9	56.8 [69]	48 [12]
Görüntü Maskeleme [38]	Yeniden oluşturma temelli	56.5	44.5	30
Renklendirme [39]	Yeniden oluşturma temelli	65.6	46.9	35.6
Görece Pozisyon Tahmini [40]	Tahmin temelli	65.3 [69]	51.1 [69]	-
Yapboz Bulmaca [41]	Tahmin temelli	67.6	53.2	37.6
Döndürme [42]	Tahmin temelli	72.9	54.4	39.1
Derin Kümeleme [43]	Kümeleme temelli	73.7	55.4	45.1

Tablo 5. *ResNet [5] mimarisi kullanılarak, Pascal VOC 2007 [67] veri setinde nesne tespiti hedef görevi performans değerlendirmesi. Nesne tespiti hedef görevinde ortalama AP değerlendirme metriği ve ResNet [5] omurgası kullanan Faster R-CNN [10] mimarisi kullanılmıştır.*

Yöntem	Kategori	Nesne Tespiti
Denetimli Öğrenme	-	81.3
PIRL [59]	Karşılaştırmalı	80.7
MoCO [26]	Karşılaştırmalı	81.5
MoCO-v2 [60]	Karşılaştırmalı	82.5
SwAV [30]	Karşılaştırmalı	82.6
SimSiam [28]	Karşılaştırmalı olmayan	82.4
Barlow Twins [29]	Karşılaştırmalı olmayan	82.6

IV. TARTIŞMA

Yeniden oluşturma temelli, tahmin temelli ve kümeleme temelli öz-denetimli öğrenme yöntemlerinin yardımcı göreve özgü temsilleri öğrenmesi, öğrenilen temsillerin hedef görevde genelleştirilememesine neden olmakta ve denetimli öğrenme ile yüksek performans farkı

oluşmaktadır. Bununla birlikte karşılaştırmalı ve karşılaştırmalı olmayan öz-denetimli öğrenme yöntemlerinin denetimli öğrenmeyle olan performans farkının azaldığı deneysel sonuçlar ile gösterilmiştir. Ancak öz denetimli öğrenme yöntemlerinin kısa sürede elde ettiği bu başarıyla birlikte bazı hususların tartışılmasına ihtiyaç duyulmaktadır.

A. HESAPLAMA GÜCÜ İHTİYACI

Öz denetimli öğrenme yöntemlerinde büyük boyutlu eğitim verisi ve yığın kullanımı nedeniyle yüksek boyutlu hesaplama gücüne ihtiyaç duyulmaktadır. Yığın boyutunun 8192 olduğu SimCLR [25] yöntemi 128 TPU v3'e sahip donanımda ve yığın boyutunun 4196 olduğu BYOL [27] yöntemi 512 TPU v3'e sahip donanımda çalıştırılmaktadır. Ayrıca öz denetimli öğrenme yöntemleri, denetimli öğrenme yöntemlerine kıyasla daha uzun eğitime ihtiyaç duymaktadır. SimCLR [25] ve BYOL [27] yöntemleri en başarılı sonuçları 1000 epoch (tüm eğitim veri setinin model üzerinden bir defa geçmesi) çalıştırılarak elde etmekteyken, denetimli öğrenme yöntemleri genellikle 100 epoch çalıştırılmaktadır. Sonuç olarak yüksek hesaplama gücü kullanmadan genelleştirilebilir temsilleri öğrenebilen öz denetimli öğrenme yöntemlerinin geliştirilmesine ihtiyaç duyulmaktadır.

B. AĞ MİMARİLERİ

Ağ mimarisinin tasarımı öz denetimli öğrenme yöntemlerinin performansını büyük ölçüde etkilemektedir [70]. Öz denetimli öğrenme yöntemlerinde ön-eğitimli ağ mimarisi olarak genellikle AlexNet [3] ve ResNet [5] omurgası kullanılmaktadır ve ön-eğitimli ağ mimarisi büyüdükçe başarımın artığı görülmüştür [61]. [71] çalışmasında öz-denetimli öğrenme yöntemlerinde kullanılan omurga çıktı boyutunun performansa etkisi incelemiş ve omurga çıktı boyutu arttıkça öz-denetimli öğrenme yöntemlerinin performansı artarken, denetimli öğrenme yöntemlerinin performansının azaldığı belirtilmiştir. Son yıllarda bilgisayarlı görü görevlerinde kullanılan diğer bir mimari de görüntü transformer mimarileridir [63]. Öz denetimli öğrenme yöntemlerinde de görüntü transformer mimarileri kullanılmış ve başarımın arttığı görülmüştür [62], [72]–[74]. Bu nedenle öz denetimli öğrenme yöntemlerinde görüntü transformer mimarileri kullanımı gelecek vadeden bir konudur.

C. VERİ ÇOĞALTMA TEKNİKLERİ

Karşılaştırmalı ve karşılaştırmalı olmayan öz denetimli öğrenme yöntemlerinde her bir görüntüye yeniden boyutlandırma, rastgele kırpma, döndürme, öteleme, renk bozma vb. veri çoğaltma teknikleri uygulanarak pozitif çiftler oluşturulmaktadır [75]. SimCLR [25] çalışmasında, veri çoğaltma teknikleri tek başlarına veya iki veri çoğaltma tekniğini birlikte uygulanarak oluşturulan çoğaltılmış görüntülerle performans değerlendirmesi yapılmıştır. Bu çalışmada birlikte uygulanan rastgele kırpma ve renk bozma veri çoğaltma yöntemlerinin yüksek başarım elde ettiği görülmüştür. SwAV [30] yönteminde ise çoklu ölçekli kırpma veri çoğaltma tekniği uygulanmaktadır. Çoklu ölçekli kırpma veri çoğaltma tekniğinde, bir görüntüden küçük boyutlarda çoklu kırpma yapılarak elde edilen görüntülere veri çoğaltma teknikleri uygulanarak çoklu çoğaltılmış görüntüler oluşturulmaktadır. Bu sayede diğer öz denetimli öğrenme yöntemlerinden farklı olarak bir görüntü için ikiden fazla çoğaltılmış görüntü kullanılırken, görüntülerin küçük boyutları sayesinde yüksek hesaplama ihtiyacından kaçınılmaktadır. SimCLR [25] yönteminde yüksek hesaplama gücüne ihtiyaç duyan birçok veri çoğaltma tekniğini kullanılırken, [76] çalışmasında veri çoğaltma yöntemi olarak sadece basit gauss gürültüsü kullanılarak genelleştirilebilir görüntü temsili elde edilebileceği gösterilmiştir ve [77] çalışmasında karmaşık veri çoğaltma tekniklerinin kullanımının modelin eğitiminde önemli ölçüde yavaşlamaya neden olacağı belirtilmiştir. Ayrıca doğal görüntülerde yaygın olarak kullanılan birçok veri çoğaltma tekniği medikal görüntüler için uygun olmadığı görülmüştür [78], [79]. Görüldüğü üzere kaliteli görüntü temsillerinin öğreniminde doğru veri çoğaltma tekniklerinin seçimi büyük öneme sahiptir [80], [81].

D. NEGATİF ÖRNEK SEÇİMİ

Karşılaştırmalı öğrenme yöntemleri, pozitif örnekleri negatif örneklerden ayırt ederek eğitilmektedir. Ancak genelleştirilebilir görüntü temsili elde edilebilmesi için çok sayıda negatif örneğe ihtiyaç duyulmaktadır [52]. Çok sayıda negatif örneği saklamak için IntDisc [50] ve PIRL [59] yöntemlerinde bellek bankası, MoCO [26] yönteminde kuyruk veri yapısı ve SimCLR [25] yönteminde ise yüksek boyutlu yığın kullanılmaktadır. Bununla birlikte pozitif örneklerden ayırt edilmesi zor olan negatif örneklerin seçilmesiyle karşılaştırmalı öğrenme yöntemlerinin daha hızlı ve iyi eğitilmesi sağlanabilmektedir [82], [83]. Bu nedenle karşılaştırmalı kayıp fonksiyonun minimize edilmesinde negatif örnek seçimi önemli bir konudur.

E. BAŞARISIZ ÇÖZÜM SORUNU

Başarısız çözüm, modelin sadece pozitif çift kullanarak eğitilmesiyle tüm girdi görüntüleri için sabit bir vektör üretmesi sorunudur. Karşılaştırmalı öğrenme yöntemlerinde başarısız çözümden kaçınmak için çok sayıda negatif çift kullanılmaktadır. Ancak son yıllarda negatif çift kullanmadan başarısız çözüm sorunu yaşamayan BYOL [27], SimSiam [28] ve Barlow Twins [29] karşılaştırmalı olmayan öz-denetimli öğrenme yöntemleri önerilmiştir. Barlow Twins [29] yöntemi farklı bir kayıp fonksiyonu kullanarak başarısız çözümü önleyebilmekteyken, BYOL [27] ve SimSiam [28] yöntemleri, başarısız çözümü önlemek için tahmin edici kafasına ve gradientin durdurulması işlemine ihtiyaç duymaktadır. Ancak negatif örnek kullanmadan genelleştirilebilir görüntü temsilleri elde edebilen bu yöntemlerin arkasındaki temel teoriyle ilgili az sayıda çalışma bulunmaktadır [84]. Bu nedenle karşılaştırmalı olmayan öz-denetimli öğrenme yöntemlerinin teorik temelleri araştırılması gereken bir konudur.

V. SONUÇ

Ayırt edici görüntü temsillerinin elde edilmesinde denetimli öğrenme yöntemleri büyük miktarda etiketli veriye ihtiyaç duymaktadır. Etiketli veri elde etmenin zaman alıcı, pahalı ve emek yoğun bir işlem olması ve etiketsiz büyük boyutlu verinin varlığı, araştırmacıların son yıllarda öz-denetimli öğrenme yaklaşımına ilgi duymalarına neden olmuştur. Etiketsiz büyük boyutlu veriden faydalanarak genelleştirilebilir görüntü temsilleri elde edebilen öz-denetimli öğrenme yöntemleri, denetimli öğrenme yöntemleriyle rekabet edecek sonuçlar elde etmişlerdir. Bu çalışmada, bilgisayarlı görü görevlerinde kullanılan öz denetimli öğrenme yöntemlerinin kapsamlı bir literatür taraması yapılarak öz denetimli öğrenme yöntemleri kategorize edilmiştir. Ayrıca öz denetimli öğrenme yöntemlerinin farklı veri setlerindeki ve görüntü sınıflandırma, nesne tespiti ve görüntü bölütleme bilgisayarlı görü görevlerindeki performans karşılaştırılması sunulmuştur. Son olarak, mevcut yöntemlerdeki sorunlu hususlar tartışılmakta ve öz-denetimli öğrenme yöntemlerinin gelecekteki potansiyel araştırma yönleri önerilmektedir.

VI. KAYNAKLAR

[1] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 2818–2826.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *in International Conference on Machine Learning (ICML*), 2019, pp. 6105–6114.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *in Advances in Neural Information Processing Systems,* 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *in International Conference on Learning Representations (ICLR)*, 2015, pp. 1–13.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] C. Szegedy et al., “Going deeper with convolutions,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[8] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv preprint arXiv:1804.02767, 2018.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *in Advances in Neural Information Processing Systems*, 2017, pp. 91–99.

[11] R. Girshick, “Fast R-CNN,” *in IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[12] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” *in IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[16] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[17] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting Unreasonable Effectiveness of Data in Deep Learning Era,” *in IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 843–852.

[18] A. V Joshi, “Amazon’s Machine Learning Toolkit: Sagemaker,” *in Machine Learning and Artificial Intelligence*, 2020, pp. 233–243.

[19] A. Chowdhury, J. Rosenthal, J. Waring, and R. Umeton, “Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations,” *Informatics*, vol. 8, no. 3, p. 59, 2021.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1717–1724.

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” *in Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[23] S. Shurrab and R. Duwairi, “Self-supervised learning methods and applications in medical imaging analysis: a survey,” *PeerJ Comput. Sci.*, vol. 8, p. e1045, 2022.

[24] A. Tendle and M. R. Hasan, “A study of the generalizability of self-supervised representations,” *Mach. Learn. with Appl.*, vol. 6, p. 100124, 2021.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” *in International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[27] J.-B. Grill et al., “Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning,” *in Advances in Neural Information Processing Systems*, 2020, pp. 21271–21284.

[28] X. Chen and K. He, “Exploring Simple Siamese Representation Learning,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15750–15758.

[29] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-Supervised Learning via Redundancy Reduction,” *in International Conference on Machine Learning (ICML)*, 2021, pp. 12310–12320.

[30] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments,” *in Advances in Neural Information Processing Systems*, 2020, pp. 9912–9924.

[31] X. Liu et al., “Self-supervised Learning: Generative or Contrastive,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857-876, 2021.

[32] K. Ohri and M. Kumar, “Review on self-supervised image recognition using deep neural networks,” *Knowledge-Based Syst.*, vol. 224, p. 107090, 2021.

[33] L. Jing and Y. Tian, “Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, 2021.

[34] Y. Bastanlar and S. Orhan, “Self-Supervised Contrastive Representation Learning in Computer Vision,” *in Applied Intelligence- Annual Volume 2022 [Working Title]*, London, United Kingdom: IntechOpen, 2022.

[35] R. Zhang, P. Isola, A. A. Efros, and B. A. Research, “Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1058–1067.

[36] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, “Self-Supervised Representation Learning: Introduction, advances, and challenges,” *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, 2022.

[37] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” *in International Conference on Machine Learning (ICML)*, 2008, pp. 1096–1103.

[38] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context Encoders: Feature Learning by Inpainting,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2536–2544.

[39] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” *in European Conference on Computer Vision (ECCV)*, 2016, pp. 649–666.

[40] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised Visual Representation Learning by Context Prediction,” *in IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430.

[41] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” *in European Conference on Computer Vision (ECCV)*, 2016, pp. 69–84.

[42] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” *in International Conference on Learning Representations (ICLR)*, 2018.

[43] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” *in European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[44] Y. M. Asano, C. Rupprecht, and A. Vedaldi, “Self-labelling via simultaneous clustering and representation learning,” arXiv preprint arXiv:1911.05371, 2019.

[45] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” *in Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[46] R. Epstein. (2023, Aug. 11). *The empty brain* [Online]. Available: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer.

[47] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1735–1742.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[49] K. Sohn, “Improved deep metric learning with multi-class N-pair loss objective,” *in Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.

[50] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised Feature Learning via Non-parametric Instance Discrimination,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[51] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” *in International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 297–304.

[52] A. van den Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” arXiv Preprint arXiv:1807.03748, 2018.

[53] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” *IEEE Trans. Pattern Anal. Mach. Intell., vol.* 38, no. 9, pp. 1734–1747, 2016.

[54] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *in International Conference on Machine Learning (ICML)*, 2016, pp. 2611–2620.

[55] O. J. Henaff et al., “Data-Efficient image recognition with contrastive predictive coding,” *in International Conference on Machine Learning (ICML)*, 2020, pp. 4182–4192.

[56] R. Devon Hjelm et al., “Learning deep representations by mutual information estimation and maximization,” *in International Conference on Learning Representations (ICLR)*, 2019, pp. 1–24.

[57] P. Bachman, R. Devon Hjelm, and W. Buchwalter, “Learning representations by maximizing mutual information across views,” *in Advances in Neural Information Processing Systems*, 2019, pp. 15535–15545.

[58] Y. Tian, D. Krishnan, and P. Isola, “Contrastive Multiview Coding,” *in European Conference on Computer Vision (ECCV)*, 2020, pp. 776–794.

[59] I. Misra and L. van der Maaten, “Self-Supervised Learning of Pretext-Invariant Representations,” *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6707–6717.

[60] X. Chen, H. Fan, R. Girshick, and K. He, “Improved Baselines with Momentum Contrastive Learning,” arXiv preprint arXiv:2003.04297, 2020.

[61] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, “Big self-supervised models are strong semi-supervised learners,” *in Advances in Neural Information Processing Systems*, 2020, pp. 22243–22255.

[62] M. Caron et al., “Emerging Properties in Self-Supervised Vision Transformers,” *in IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

[63] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929, 2020.

[64] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” arXiv preprint arXiv:2304.07193, 2023.

[65] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, “Scaling and benchmarking self-supervised visual representation learning,” *in IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6390–6399.

[66] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” *in Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[67] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[68] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes Challenge: A Retrospective,” *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[69] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell, “Data-dependent initializations of convolutional neural networks,” arXiv preprint arXiv:1511.06856, 2015.

[70] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, “A theoretical analysis of contrastive unsupervised representation learning,” *in International Conference on Machine Learning (ICML)*, 2019, pp. 9904–9923.

[71] F. Bordes, S. Lavoie, R. Balestriero, N. Ballas, and P. Vincent, “A surprisingly simple technique to control the pretraining bias for better transfer: Expand or Narrow your representation,” arXiv preprint arXiv:2304.05369, 2023.

[72] Z. Xie et al., “Self-Supervised Learning with Swin Transformers,” arXiv preprint arXiv:2105.04553, 2021.

[73] X. Chen, S. Xie, and K. He, “An Empirical Study of Training Self-Supervised Vision Transformers,” *in IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 9620–9629.

[74] C. Li et al., “Efficient Self-supervised Vision Transformers for Representation Learning,” arXiv preprint arXiv:2106.09785, 2022.

[75] S. Albelwi, “Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging,” *Entropy*, vol. 24, no. 4, p. 551, 2022.

[76] F. Bordes, R. Balestriero, and P. Vincent, “Towards Democratizing Joint-Embedding Self-Supervised Learning,” arXiv preprint arXiv:2303.01986, 2023.

[77] R. Balestriero et al., “A Cookbook of Self-Supervised Learning,” arXiv preprint arXiv:2304.12210, 2023.

[78] C. Zhang, Z. Hao, and Y. Gu, “Dive into the Details of Self-Supervised Learning for Medical Image Analysis,” *Med. Image Anal.*, vol. 89, p. 102879, 2023.

[79] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, “Self-supervised learning for medical image classification: a systematic review and implementation guidelines,” *NPJ Digit. Med.*, vol. 6, no. 1, p. 74, 2023.

[80] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?,” *in Advances in Neural Information Processing Systems*, 2020, pp. 6827–6839.

[81] X. Wang and G. J. Qi, “Contrastive Learning with Stronger Augmentations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, 2022.

[82] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, “Hard negative mixing for contrastive learning,” *in Advances in Neural Information Processing Systems*, 2020, pp. 21798–21809.

[83] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, “Contrastive Learning with Hard Negative Samples,” arXiv preprint arXiv:2010.04592, 2020.

[84] Y. Tian, X. Chen, and S. Ganguli, “Understanding self-supervised Learning Dynamics without Contrastive Pairs,” *in International Conference on Machine Learning (ICML)*, 2021, pp. 10268–10278.