Universidad del País Vasco
Euskal Herriko Unibertsitatea
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA

Gradu Amaierako Lana / Trabajo Fin de Grado

Fisikako Gradua / Grado en Física

Ikasketa sakona erabiliz fisika intuitiboki ikasi dezakeen ikusmen artifizialeko sistema baten diseinua

Jon Perez Visaires

Zuzendaria:
Mikel Peñagarikano Badiola

Elektrizitatea eta Elektronika Saila
Zientzia eta Teknologia Fakultatea
Euskal Herriko Unibertsitatea UPV/EHU

Leioan, 2020ko irailean

Gaien Aurkibidea

1 Sarrera eta helburua

1.1 Sarrera

Ikustea, kontzeptualki zer dagoen eta hori espazialki kokatua non dagoen begiratuz jakitea da, baina etorkizun hurbilean gertatuko dena aurresatea eta horretan oinarrituz hartu daitezkeen erabakiak igartzea ikustea da ere [15]. Bestalde, ikustea, elementu bisualen kokapena detektatzeaz gain, hauek dituzten ezaugarri fisikoak eta haien arteko harremanak antzematea da [1]. Errealitatearen osagaiak diren elementuek jarraitzen dituzten lege fisikoei buruz arrazoitzeko ahalmena giza-adimenaren funtsezko oinarria da eta, ondorioz, adimen artifizialaren (*Artificial Intelligence*, AI) helburu garrantzitsuenetarikoa.

Adimen artifizialaren jakintza-arlo zabalaren barnean ordenagailu bidezko ikusmen artifiziala (*Computer Vision*, CV) oinarrizko gaia da, baina irudi eta bideoekin lan egitea askotan ez da eginkizun erraza: irudi bakar batean kodetua dagoen informazio-kopurua itzela da eta (pixel bakarrean 0 eta 255 artean dauden 3 zenbakizko balio gordetzen dira, kolorezko RGB irudien kasuan). Hortaz, ikusmenarekin lan egiten duten sistemak informazio-fluxu handia jasateko edo sarrerako informazio hori hasieran nolabait sinplifikatzeko gai izan behar dira. Gainera, adimen artifizialeko algoritmoek ikusmen-lan konplexuetan giza-gaitasunera iritsi daitezen, sistemek fenomeno ugari ulertu behar dituzte, bereziki mundu makroskopikoan gertatzen diren fenomeno fisikoak. Hortaz, ikusmen artifizialeko sistemek objektu makroskopikoak, mugimendua, indarrak eta antzeko kontzeptu fisikoak zentzu intuitibo batean ulertzeko eta inguruko errealitatea interpretatzeko ahalmena eskuratu behar dute, haiekin lortu nahi den informazioa baliagarriagoa izan dadin [22].

Heldutasunera heltzen garenean, gizakiok dagoeneko mundu fisikoari buruz jakintza sakona eskuratu dugu, bizitza osoan zehar behin eta berriz izandako esperientzien kopuru handiaren ondorio zuzena. Adibidez, badakigu objektu zurrun bat ez dela estuago den zulo batean sartuko, ezkutuan dauden objektuek existitzen jarraitzen dutela edo esku artean dugun zerbait askatzerakoan grabitatearen ondorioz lurrerantz eroriko dela [8]. Hasiera batean, bularreko umeek gertakizun fisikoei buruzko ezagutza oso txikia zutela pentsatzen zen [19], baina 1980ko hamarkadan egindako esperimentuen bidez ideia hau ezeztatu egin zen. Izan ere, ume oso txikiek espero ez dituzten gertakizun fisikoen aurrean harridura eta arreta handia adierazten dute eta, horren ondorioz, objektuen zinematikari eta dinamikari buruzko iragarpen sinpleak egiteko gaitasuna dutela frogatu zen [8]. Ikusmen artifizialeko sistemek, berriz, ez dute horrelako ulermen fisikoa eskuratzeko aukerarik izan, ahalmen hau esplizituki programatzen ez bada, behintzat.

Umeek txikitatik inguruan ikusten duten mundua modu intuitibo batean fisikoki interpretatzen ikasten dute, era guztiz autonomoan eta behin eta berriz errepikatzen diren esperientzia fisikoetan oinarrituz. Gaur egun, psikologia kognitiboaren arloan teoria nagusia gizakiok jaiotzetik gero esperientziarekin doitzen den fisika intuitiboki lantzeko balio duen barne arrazonamendu-sistema bat dugula da; kausalitatea ulertzeko sistema konputazional abstraktu bat, hain zuzen ere [8]. Hori dela eta, umeek Newtonen legeak jarraitzen dituen pilota

batekin jolasten duten bakoitzean, errealitatea gobernatzen duten lege fisiko hauen portaeran sakontzeko gai dira. Beraz, ume txikiek objektu makroskopikoek espazioan eta denboran zehar dituzten elkarrekintzak ulertzeko eta haien ibilbideak jarraitzeko gaitasuna oso azkar eskuratzen dute. Haurtzaroan zehar dituzten esperientzia guzti hauetatik jakintza fisikoa lortzen dute eta esperientzien errepikapenaren ondorioz ikasteko gai dira, gaur egun interes handia lortu duten ikasketa sakoneko (*Deep Learning*, DL) neurona-sareen (*Artificial Neural Network*, ANN) ikasketa-prozesuaren antzekoa den metodoa.

Azken urte hauetan ikusmen artifizialaren esparruan eman diren aurrerapausoak izugarriak izan dira, baina adimen artifizialeko sistemak oraindik gehienetan motz gelditzen dira nahiko konplexuak diren eszena bisualak interpretatzerako orduan, giza-errendimenduarekin konparatzen baditugu behintzat [22]. Objektu makroskopikoen zinematikari eta dinamikari buruzko ulermen bat lortzea ez da eginkizun erraza, batez ere eremu horiek gobernatzen dituzten lege fisikoak kasu desberdin askorako eskuz programatu behar badira. Oztopo hau gainditzeko, ikasketa automatikoko sistemek era autonomoan ikasteko duten gaitasuna erabiltzea irtenbide bat izan daiteke. Hori dela eta, fisika intuitiboa nolabait ulertu dezakeen ikasketa sakoneko neurona-sareetan oinarritutako ikusmen artifizialeko sistema bat proposatzen da lan honetan.

1.2 Helburua

Zeregin bisual konplexuetan ikusmen artifizialeko sistemek giza-errendimendua lortzeko, errealitatean objektu makroskopikoek jarraitzen dituzten lege fisikoak ulertu behar dituzte, edo gutxienez lege hauei buruzko intuizio sinple bat garatu behar dute. Nahiz eta ulermen fisikoa ikuste-esperientzietatik jakintza lortzeko ezinbesteko tresna izan, gaur egun ikusmen-lanetarako erabiltzen diren ordenagailu bidezko ikusmen artifizialeko sistema askok ez dute ulermen mota hau gehiegi kontuan hartzen [22].

Umeek era autonomoan informazio fisikoa intuitiboki ulertzeko eta ikasteko duten gaitasunean oinarrituz, adimen artifizialeko aplikazio sendoagoak eta moldakorragoak eraikitzeko ikusmen-sistemek objektuen zinematikari eta dinamikari buruzko informazioa eskuratzea eta hura interpretatzeko ahalmena izatea ezinbestekoa da. Izan ere, etorkizuneko gertakizunak aurresateko gaitasuna giza-adimenaren funtsezko osagaia eta sistema adimentsu askoren oinarria da, eta gaitasun hau denbora errealeko adimen artifizialeko sistemek (ibilgailu autonomoak, lan-edo etxe-robotak etab.) erabaki egokiak momentuan hartu ahal izateko behar-beharrezkoa da [17].

Bideo batean etorkizuneko fotogramak zehaztasunez aurresaten ikasi ahal izateko, adimen artifizialeko sistema batek bideoa osatzen duten irudien eduki kontzeptuala eta fotograma hauetan agertzen diren elementu bisualen zinematika eta dinamika derrigorrez maila batean modelizatu behar ditu, eta horretarako inguruko mundu fisikoaren barne-errepresentazio zehatza eta ez-tribiala garatzea beharrezkoa da [16]. Beraz, bideo baten etorkizuneko fotogramak eraginkortasunez iragarri ahal dituen sistema fisika intuitiboki nolabait ikasi duen sistema da ere.

Hau guztia kontuan hartuz, lan honen helburu nagusia oinarrizko intuizio fisikoa lortzeko gaitasuna duen ikusmen artifizialeko sistema bat garatzea da, horretarako neurona-sareak eta ikasketa sakonaren arloko teknikak erabiliz. Sistema hau objektu makroskopikoen zinematikaren eta dinamikaren hurbilketa bat egiteko gai izatea lortu nahi da, eta horretarako hainbat mota desberdineko objektuak mugimenduan adierazten dituzten bideoetan etorkizuneko fotograma bat iragartzea helburu bezala duen neurona-sare bat diseinatu, inplementatu eta entrenatuko da.

2 Adimen artifiziala

Azkenengo urteetan, adimen artifiziala (*Artificial Intelligence*, AI) komunikabideetan asko landu den gaia izan da. Ikasketa automatikoa, ikasketa sakona eta adimen artifiziala maiz agertzen dira gaur egun prentsa artikuluetan, askotan teknologian espezializatuak ez diren publikazioetan ere. Hauetan, etorkizun hurbilean gizakiok gaur egun egiten ditugun lan asko makinek egingo dituztela aipatzen da, baita ekintza ekonomiko gehienak botek edo agente adimentsuek burutuko dituztela. Izan ere, kasu batzuetan adimen artifiziala lanmundura ailegatu da jada, eta gaur egun langile ugarik algoritmo batek era autonomoan hartutako erabakiak onartu behar dituzte; pertsona askoren benetako nagusia prozesu zehatzak optimizatzeko programatua dagoen adimen artifizialeko sistema bat da [4]. Horrela, software adimentsua lan-errutinak automatizatzeko, irudiak edo hizkera ulertzeko eta medikuntzan diagnostikoak egiteko erabiltzen dugu jada [6]. Hala ere, garrantzi handia du adimen artifizialak benetan lortu dituen arrakastak eta lorpenak zeintzuk diren ezagutzeak, ahalmenak eta mugak ezagutuz errazagoa izango baita teknologia hau era egokian aplikatzea [3].

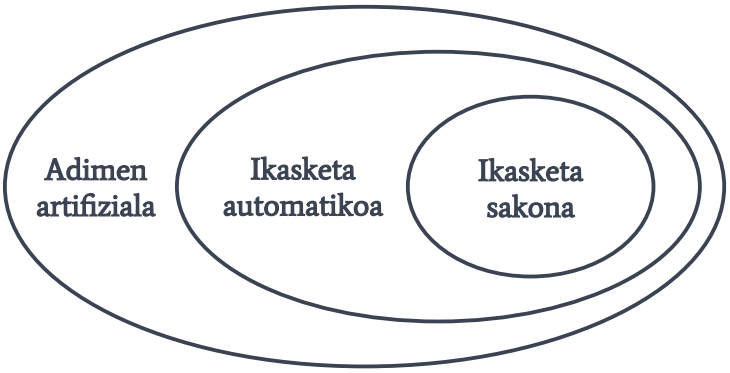
Adimen artifiziala 1950ko hamarkadan sortu zen informatikaren azpi-arloa da, eta agente adimentsuen diseinua eta portaera aztertzen dituen jakintzagaia da [21]. Garai hartan, informatika jaio berria zen ere esparru teoriko bezala, eta hainbat aurrendarik hurrengo galdera erantzun nahi izan zuten: ordenagailu batek adimena garatu dezake? Galdera honen ondorio sakonak gaur egun hausnartzen jarraitzen ditugu [3]. Adimen artifizialaren helburu zientifiko nagusia sistema artifizialetan edo naturaletan portaera adimentsua posiblea egiten duten printzipioak ulertzea da, horretarako adimena konputazioan oinarritua dagoela hipotesi bezala hartuz [21].

Agente bat bere ingurunean eragina izan dezakeen zerbait da. Agente adimentsu bat, berriz, ingurunean eragina izateko ahalmen hori helburu jakin bat lortzeko eta momentuko baldintza zehatzak kontuan izanda modu adimentsu batean erabiltzen duen agentea da [20]. Gainera, agente bat adimentsua izan dadin ingurune eta helburu aldakorretara moldatzeko, esperientziatik ikasteko eta bere pertzepzio- eta konputazio-limiteen barnean erabaki zentzudunak hartzeko gaitasuna izan behar du [21]. Lan honetan proposatzen den ikusmen artifizialeko sistema mota honetako agente adimentsuek duten ingurunearen pertzepzioa hobetzeko modu bat izan daiteke.

Adimen artifizialaren beste definizio labur bat hurrengoa izan daiteke: normalean gizakiok burutzen ditugun zeregin intelektualak automatizatzeko ahalegina [3]. Definizio hau kontuan hartuz, ikasketa automatikoa eta ikasketa sakona adimen artifizialaren arlo orokorraren barnean sartzen dira (2.1 Irudia), baina ikasketa-prozesurik erabiltzen ez dituzten adimen artifizialaren beste hainbat azpi-arlo ere existitzen dira. Adibidez, hasiera batean sortu ziren xakean jolasteko programak era esplizituan idatzitako erregela-zerrenda luzeez baliatzen ziren [25], eta ez zituzten ikasketa-prozesu automatikorik erabiltzen (ez dira ikasketa automatikoaren eremuan sartzen).

Denbora luzez, adimen artifizialaren arloko aditu askok informazioa eraldatzeko erregela esplizituen zerrenda behar bezain luzea sortuz giza-mailako adimena lortu ahal zela pentsatzen zuten [3]. Planteamendu hau adimen artifizial sinboliko izenaz ezagutzen da eta 1950ko hamarkadatik 1980ko hamarkadaren amaieraraino paradigma nagusia izan zen, batez ere 1980ko hamarkadan agertu ziren sistema adituen garaian [20]. Horrela, hasiera batean gizakiontzat intelektualki zailak ziren problemak sistema aditu hauen bidez azkar ebatzi ziren, ondo definitutako erregela logikoak jarraitzen dituzten zereginak ordenagailuentzat ebazteko errazak ziren eta [6].

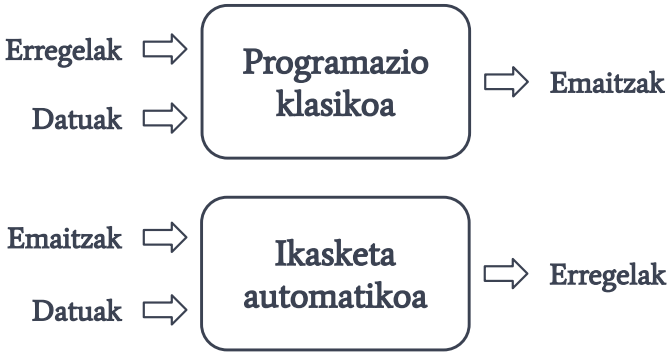
Hala ere, sistema adituek gizakiok ia automatikoki burutzen ditugun zereginetan zailtasunak zituzten, batez ere giza-pertzepzioarekin (ikusmena, entzumena etab.) lotutako problemetan. Arazo nagusia mota desberdineko jakinduria kodetzeko baliagarriak diren erregelak gizakiek logikoki definitu eta esplizituki programatu behar zituztela zen, eta gizakiok automatikoki eta intuitiboki burutzen ditugun zeregin kognitiboak formalki definitzea oso konplexua izan daiteke [27][6][1]. Adimen artifizial sinbolikoak era egokian ebazten zituen ondo definitutako problema logikoak (lehen aipatutako xakean jolasteko gaitasuna [25], adibidez), baina problema kognitibo konplexuagoak ebazteko behar ziren erregela esplizitu guztiak eskuz definitzea ezinezkoa zela argi geratu zen denborarekin [3]. Problema zail hauen adibide batzuk irudi-sailkapena, irudi batean aurpegiak detektatzea, hizkuntza itzultzaile automatikoak eta hizketa-ezagutzea dira [6]. Arazo honi konponbidea emateko, adimen artifizialaren azpi-arlo berri bat agertu zen, ikasketa automatikoa.



Irudia 2.1: Adimen artifizialaren, ikasketa automatikoaren eta ikasketa sakonaren arteko erlazioa adierazten duen Venn diagrama.

2.1 Ikasketa automatikoa

Ikasketa automatikoaren (*Machine Learning*, ML) eremua hurrengo galdera erantzuten saiatzean sortu zen: posiblea ahal da ordenagailu batek zeregin zehatz bat bere kabuz burutzen ikastea? Programatzaileek eskuz idatzitako datu-prozesamendurako erregelak erabili beharrean, ordenagailu batek modu guztiz autonomoan eta bakarrik eskuragarri dituen datuak erabiliz informazioa era baliagarri batean eraldatzen dituzten erregela logikoak automatikoki ikasteko gai izan ahal da? [3] Beraz, ikasketa automatikoa ereduen ezagutzean eta datuetatik automatikoki ikastean oinarrituta dagoen adimen artifizialaren azpi-arloa da [24].



Irudia 2.2: Programazio klasikoaren eta ikasketa automatikoaren paradigmen alderaketa.

Programazioaren paradigma klasikoan, adimen artifizial sinbolikoan esaterako, erregelak (programa) eta prozesatu beharreko datuak sarrera moduan ezartzen dira, eta irteeran ordenagailuak bueltatutako erantzunak lortzen dira. Ikasketa automatikoa erabiltzean, ordea, sarrera bezala datuak eta lortu nahi diren erantzunen adibideak erabiltzen dira, eta irteeran ordenagailuak sarreran eman zaizkion erantzun hauetara hasierako datuetatik heltzeko behar dituen erregela logikoak lortzea espero da (2.2 Irudia) [3]. Prozesu honen bidez lortutako programazio-erregelak ordenagailuak inoiz ikusi ez dituen datu berrietan aplikatu daitezke, horrela zeregin zehatz bat ebazteko balio duten erantzun berriak lortuz, algoritmoaren orokortze-ahalmena ona bada.

Hori dela eta, ikasketa automatikoko sistema bat esplizituki programatu beharrean entrenatzen dela esan ohi da. Sistemari ebatzi behar duen zereginarekin lotuta dauden adibide asko aurkezten zaizkio (sarrera-irteera bikoteen laginak), eta adibide hauetan sistemak zeregina modu automatikoan ebazteko behar dituen erregelak definitzea ahalbidetzen dioen egitura estatistiko inplizitua aurkitzen saiatzen da [3]. Adibide moduan, duela urte batzuk sare sozialetan eskuz etiketatu behar ziren argazkiak gaur egun era guztiz automatikoan zeregin bera betetzen duten ikasketa automatikoko sistemak entrenatzeko erabiliak izan dira. Hortaz, sistema hauek irudietan agertzen diren aurpegiak etiketa eta pertsona zehatzekin lotzeko beharrezkoak diren programazio-erregelak definitzeko gaitasuna garatu dute [18]. Aurreko guztia kontuan hartuz, ikasketa automatikoa erabiltzeko hiru gauza behar ditugu [3]:

- Sarrerako datuak: zeregina irudiak etiketatzea bada, sarrerako datuak irudiak izango dira.
- Itxarondako irteerak: sarrerako irudi bakoitzarekin lotuta dauden etiketak izango dira; "txakurra", "katua" edo "pertsona", esaterako.
- Algoritmoaren eraginkortasuna neurtzeko modu bat: azkenengo puntu hau beharbeharrezkoa da algoritmoaren irteeraren eta itxarondako irteeraren arteko diferentzia kalkulatzeko. Neurketa hau berrelikadura-seinale bezala erabiltzen da algoritmoak lan egiteko duen modua aldatzeko. Azkenengo pausu honetan egiten den doikuntza honi *ikasketa* deritzogu.

Hortaz, ikasketa automatikoko sistema batek sarrerako datuak irteera esanguratsuetan bihurtzen ditu, horretarako lehenengoz algoritmoa entrenatzeko eskuragai dauden sarrera eta irteera bikoteen adibide ugari behatuz eta datu horietatik informazio estatistikoa automatikoki eskuratuz [6]. Horrela, ikasketa automatikoaren helburu nagusia datuak era esanguratsu batean transformatzea da, hau da, sarrerako datuetatik abiatuz irteera adierazgarriak sortzeko gaitasuna garatzea. Sistema hauek onak badira, haien irteerak eta itxarondako benetako irteerak antzekoak izango dira. Izan ere, errepresentazio jakin batean zailak diren problema batzuk asko errazten dira modu egokian adierazten badira, fisikan problema mota zehatz batzuk ebazteko koordenatu kartesiarretatik polarretara aldaketa egiten denean gertatzen den bezala. Beraz, ikasketa automatikoaren testuinguruan, *ikasketa* zeregin zehatz bat ebazteko sarrerako datuen errepresentazio egokiaren bilaketa automatikoaren prozesua da, horretarako sarrerako datuen errepresentazio hauen egokitasuna neurtzen duen berrelikadura-seinale bat erabiliz [3][6].

Ikasketa automatikoaren arloak matematika estatistikoarekin lotura estua du, baina haien artean desberdintasun nabariak daude ere. Estatistikan gertatzen ez den bezala, sistema hauek datu-multzo handi eta konplexuekin lan egiten dute gehienetan: milioika irudi eta irudi bakoitzaren barnean milaka pixel dituzten datu-multzoak, esaterako. Irudiak, bideoak, grabazioak eta antzekoak datu-multzo ez-egituratu hauen adibideak dira. Kasu konplexu hauetan, analisi estatistiko klasikoa aplikatzea oso zaila edo zuzenean ezinezkoa izango litzateke [3]. Gainera, ikasketa automatikoaren eta ikasketa sakonaren arloetan agertzen den teoria matematikoa ez da oso zabala, eta gaur egun askotan ideiak teorikoki baino enpirikoki frogatzen dira.

2.2 Ikasketa sakona

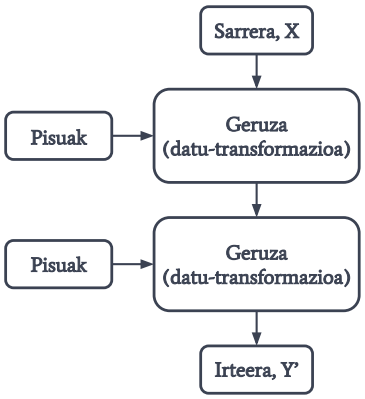
Ikasketa sakona (*Deep Learning*, DL), batzuetan errepresentazio-ikasketa hierarkikoa edo errepresentazio-ikasketa geruzatua ere deitua [3], ikasketa automatikoaren barnean dagoen azpi-arlo bat da (2.1 Irudia). Ikasketa sakonak lehen aipatu diren giza-pertzepzioarekin lotuta dauden problema konplexuak hobeto ebazteko modua eskaintzen du. Mota honetako algoritmoen bidez, ordenagailuek esperientziatik ikasteko eta mundua kontzeptu-hierarkia

baten moduan ulertzeko ahalmena eskuratzen dute. Kontzeptu-hierarkia honi esker, ordenagailuek kontzeptu konplexuak sinpleagoak diren beste hainbat kontzeptu konbinatuz ikastea posiblea da [6]. Gainera, esperientziatik era autonomoan ikasten dutenez, beste sistema mota batzuetan ezinbestekoa den gizakiok formalki definitutako erregela-zerrendaren beharrik ez dute.

Ikasketa sakoneko arkitekturek ondoz ondoko geruza (datu-transformazio sinpleak) ugari erabiltzen dituzte, sarrerako datuen errepresentazio gero eta adierazgarriagoak eraikitzeko ahalmena garatu ahal izateko [3]. Tradizionalki, ohikoena bat edo gehienez bi geruza ezkutu zituzten arkitekturekin lan egitea zen (geruza ezkutuak sistemaren sarreran edo irteeran ez dauden guztiak dira). Sareak diseinatzeko modu honen atzean zegoen arrazoi nagusia sakonagoak ziren neurona-sareak entrenatzea oso zaila zela zen, une hartan sareak entrenatzeko existitzen ziren metodoak eta algoritmoak eraginkortasun txikikoak eta nahiko konplexuak zirelako. Neurona-sare hauek barne-errepresentazio abstraktuak sortzeko zailtasunak zituzten, ikasketa sakoneko algoritmoen abstrakzio-ahalmena sarearen sakontasunarekin eta geruzen kopuruarekin zuzenean lotuta dago eta [6]. Hala eta guztiz ere, mende honetan egindako aurrerapenei esker, gaur egun ehunka geruza dituzten neurona-sareak entrenatzeko gaitasuna eskuratu dugu. Hori dela eta, ikasketa automatikoaren esparruan garrantzitsuak diren problema asko ebazterako orduan ikasketa sakoneko arkitekturak nagusi dira.

Neurona-sareek, ikasketa automatikoko beste algoritmoekin konparatuz, errepresentazio hobeak sortzen dituzte, geruza bakoitzak aurreko geruzen irteeretan oinarrituta sarrerako datuen errepresentazio gero eta abstraktuago bat sortzen duelako; sakonera eta, ondorioz, geruzen kopurua handitzeak sareak sortu ahal duen barne-errepresentazioaren konplexutasuna eta abstrakzio-maila handitzen ditu ere. Horrela, sarearen hasierako geruzek behe-mailako informazioa eskuratzen dute, eta behe-mailako informazio honen bidez hurrengo geruzek sarrerako datuen errepresentazio hobeagoa sortu dezakete, sareak ebatzi behar duen problemaren soluzioa aurkitzeko baliagarriak diren ezaugarriak biltzen dituena [3]. Errepresentazio hauek gehienetan ez dira gizakiontzat oso argiak edo ulergarriak, baina neurona-sarearen barnean haien helburua betetzen dute [6]. Horrela, neurona-sare bat informazioa destilatzen duen etapa anitzeko eragiketa dela esan daiteke, non informazioa ondoz ondoko iragazki zehatz batzuetatik igaro ondoren sarearentzat gero eta adierazgarriago bihurtzen den.

Azkenengo urteetan ikasketa sakonak izan duen arrakastaren arrazoi nagusia hainbat zereginetan ikasketa automatikoaren arloko beste teknika batzuekin alderatuz errendimendu hobeagoa eskaintzen duela da, batez ere giza-pertzepzioa imitatu nahi duten sistemen kasuan. Bestalde, ikasketa sakonak hainbat problemen ebazpena asko errazten du: beste metodo batzuk erabili ahal izateko eskuz definitu behar diren datuen ezaugarriak neurona-sareek era automatiko batean ateratzen dituzte [6]. Ikasketa sakonaren esparrua ez da berria, baina gaur egungo hardware aurrerapenek, bereziki prozesatzeko unitate grafikoetan (*Graphic Processing Unit*, GPU) egindakoak, lehen egingarriak ez ziren arkitekturak entrenatzea ahalbideratu dute. Hauen artean, aurrerago azalduko diren konboluzio neurona-sareek garrantzi handia dute ikusmen artifizialaren arloan.



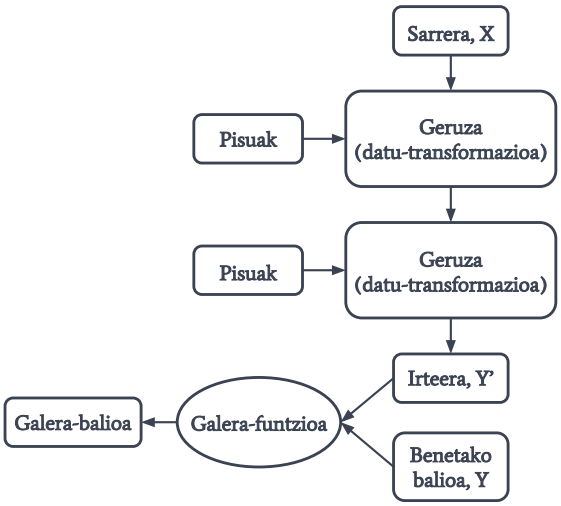
Irudia 2.3: Neurona-sarea osatzen duten geruzak haien pisuen bidez parametrizatuak daude.

2.3 Neurona-sareak

Ikasketa automatikoaren funtsezko helburua sarrerako datuak (irudiak, esaterako) itxarondako irteerekin ("txakurra" moduko etiketak edo irudiaren segmentazio-mapa, adibidez) lotzea da, horretarako entrenamendu-ziklo batean benetako sarrera-irteera bikoteen lagin asko behatuz [24]. Neurona-sare sakonek lotura hau ondoz ondoko geruzak (datu-transformazio sinpleak) erabiliz gauzatzen dute, eta geruza hauek entrenamendu-zikloan sarrera-irteera lagin-multzoak behatuz parametrizatzen dira [3]. Neurona-sareak ikasketa sakoneko sistemen funtsezko osagaiak dira eta gure burmuineko neurona-konexioetan kontzeptualki inspiratuta daude [24]. Hala ere, azpimarratu beharra dago neurona-sareak *ez* direla giza-burmuinaren modelizazio bat: neurona biologikoen portaera sinplifikatuan oinarrituta dauden algoritmoak dira.

Geruza bakoitzak bere sarrerako datuei aplikatzen dien transformazioa geruza horren pisuak zehazten dute, hau da, neurona-sarearen geruza bakoitzak burutzen duen datu-transformazioa zehatza pisuez parametrizatuta dago: askotan pisuak sarearen parametroak direla esaten da ere (2.3 Irudia). Testuinguru honetan, ikasketa-prozesua neurona-sarea osatzen duten geruza guztientzat problema zehatz bat ebazteko balio duten pisu egokienak bilatzea da, sarrerako datuak eta itxarondako irteerak lotzea ahalbidetuko duten parametroak, hain zuen ere [3]. Helburua argi dago, baina gaur egun erabili ohi diren neurona-sareek milioika pisu dituzte guztira, eta kontuan izan behar da parametro baten aldaketa beste parametroetan ere eragina izan ahal duela. Hortaz, sareen parametroak era egokian doitzeko (neurona-sareak *entrenatzeko*) mekanismo eraginkor bat beharrezkoa da.

Neurona-sare baten irteera kontrolatzeko eta bere errendimendua hobetu ahal izateko, lehenengo pausoa sarearen irteera eta itxarondako irteeraren arteko diferentzia neurtzeko gai izatea da. Lan hau sarearen *galera-funtzioaren* zeregina da, batzuetan ere helburu-funtzioa deitua. Galera-funtzioak neurona-sarearen iragarpenak eta itxarondako benetako balioak sarrera moduan hartzen ditu, eta haien arteko diferentzia-metrika bat kalkulatzen du [24]. Beraz, galera-funtzioak neurona-sarearen errendimendua neurtzeko balio du, sareak bueltatzen dituen irteeren eta irteera idealen arteko diferentzia emanez (2.4 Irudia) [3].



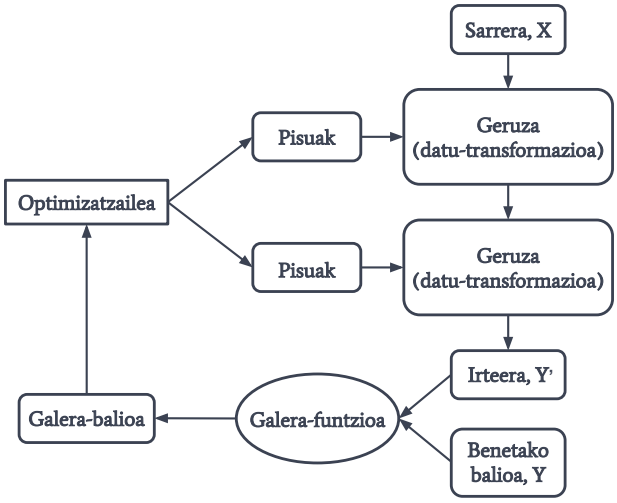
Irudia 2.4: Neurona-sarearen irteeren kalitatea neurtzeko galera-funtzio bat erabiltzea beharrezkoa da.

Galera-funtzioak bueltatutako galera-balioa berrelikadura-seinale bezala erabiltzen da sarearen parametroen balioak norabide zehatz batean doitzeko: galera-funtzioaren balioa txikitzen duen norabidean, hain zuzen ere. Doikuntza hau sarearen *optimizatzaileak* burutzen du (2.5 Irudia). Hasiera batean, sarearen parametroak ausazko balio txikiekin hasten dira eta, honen ondorioz, geruzek sarrerako datuetan zorizko datu-transformazioak aplikatzen dituzte. Hori dela eta, ohikoa da neurona-sare baten hasierako irteerak oso kalitate txarrekoak izatea: lehenengo irteera hauek itxarondako balioetatik oso urrun egongo dira. Hortaz, hasieran galera-funtzioak emango dituen neurketak handiak izango dira, sareak bueltatutako irteeren eta itxarondako balioen arteko diferentzia handia izango delako [3].

Sareak sarrera-irteera bikoteen lagin-multzo bat prozesatzen duen bakoitzean, bere parametroak norabide zuzenean poliki-poliki doitzen dira eta, ondorioz, galera-balioa txikitzen doa. Prozesu honi entrenamendu-zikloa deritzogu, eta zikloa behin eta berriz errepikatuz galera-funtzioa minimizatzen duten sarearen parametroak lortzea posiblea da [3]. Galera-funtzioaren minimoa lortzen duen neurona-sarea guztiz entrenatua dagoela esan ohi da, eta kasu horretan sarearen iragarpenak itxarondako balioetatik ahalik eta hurbilen egongo dira.

2.3.1 Gradientean oinarritutako optimizazioa

Neurona-sarearen parametroak doitzeko optimizatzaile bat erabiltzen da. Optimizazioa $f(x)$ galera-funtzio bat x parametroa aldatuz minimizatzearen zeregina bezala definitzen da [6]. Diferentziagarria den $f(x)$ funtzio baten minimoa analitikoki aurkitzea teorikoki posiblea da: $f(x)$ funtzioaren minimoa $f'(x) = 0$ betetzen duen puntu batean egon behar da derrigorrez; baldintza hau betetzen duten puntu guztiak aurkituz, $f(x)$ funtzioaren balio txikiena ematen duen puntua $f(x)$ funtzioaren minimo globala izango da. Funtzioaren parametroa tentsore bat denean, deribatuaren ordez gradientea erabili behar da, $\nabla_{\mathbf{x}} f(\mathbf{x}) = 0$.



Irudia 2.5: Galera-balioa berrelikadura-seinale moduan erabiltzen da optimizatzailearen bidez sarearen pisuak doitzeko.

Aurreko metodo analitikoa neurona-sareen kasuan aplikatzea ezinezkoa da, sareak milioika parametro izateak milioika aldagai dituen ekuazio polinomikoa ebatzi behar dela suposatzen duelako: $\nabla_{\mathbf{x}} f(\mathbf{x}) = 0$, non $\mathbf{x}$ sarearen pisuak adierazten dituen tentsorea den [3]. Oztopo hau gainditzeko, *Stochastic Gradient Descent* (SGD) izeneko algoritmoa erabili ohi da:

1. Neurona-sarea entrenatzeko eskuragai dagoen datu-multzotik ausazko lagin-sorta bat (X) eta hauekin lotuta dauden benetako irteerako balioak (Y) hartzen dira.
2. Sarrera bezala X sortaren laginak hartuz, neurona-sareak emandako irteerak lortzen dira, $\widehat{Y}$ (*forward pass*).
3. Lagin-sorta honetan sareak duen galera-funtzioaren balioaren batezbestekoa kalkulatzen da, Y eta $\widehat{Y}$ arteko diferentzia adierazten duen neurketa.
4. Galera-funtzioaren gradientea sareak momentu horretan dituen parametroekiko kalkulatzen da (*backward pass*).
5. Sarearen parametroak gradientearen kontrako norabidean apur bat doitzen dira, X lagin-sorta zehatz honetan sareak duen galera-funtzioaren balioaren batezbestekoa pixka bat txikituz. $\widehat{\mathbf{x}}$ pisu berrien tentsorea, $\mathbf{x}$ pisu zaharren tentsorea, ϵ ikasketa-erritmoa eta $f(\mathbf{x})$ galera-funtzioa badira,

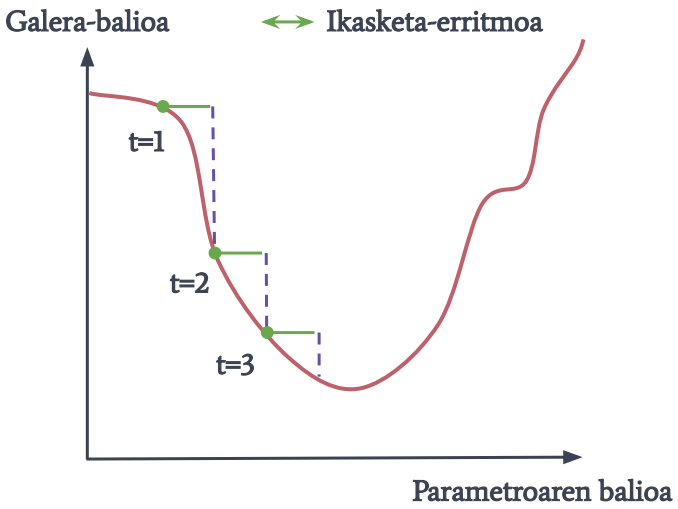
$$\widehat{\mathbf{x}} = \mathbf{x} - \epsilon \cdot \nabla_{\mathbf{x}} f(\mathbf{x}) \tag{2.1}$$

Stochastic Gradient Descent prozesu estokastikoa dela esaten da, aurreko algoritmoan erabiltzen diren lagin-sortak datu-multzotik ausazko moduan aukeratzen direlako, optimizazio-prozesuan datu-multzo osoa erabili beharrean [3]. SGD optimizazio-algoritmoa datu-multzo

handiekin lan egiteko metodo erabiliena da: nahiz eta batzuetan galera-funtzioaren minimo batera ez heldu, sarea zeregin jakin batean erabilgarria izateko nahikoa den galera-funtzioaren balio batera azkar iritsi ohi da [6].

Bestalde, neurona-sare bat entrenatzerako orduan ikasketa-erritmoaren balio egokia aukeratzeak garrantzi handia du. Ikasketa-erritmoa txikiegia bada, galera-funtzioaren gradientearen kurban egin behar den beherapenak iterazio gehiegi beharko ditu edo minimo lokal batean trabatuta geratu ahal da. Ikasketa-erritmoa handiegia bada, ordea, kurban zehar ematen diren pausoak gehiegizkoak izan ahal dira eta minimo globala atzean geratu daiteke [24].

2.6 Irudian dimentsio bakarreko adibidea adierazten da, parametro bakarra duen neurona-sarea lagin bakar batekin entrenatzen denean gertatzen dena. Errealitatean gertatzen den prozesua konplexuagoa da eta SGD algoritmoa dimentsio ugariko espazioetan aplikatzen da: neurona-sare baten parametro bakoitza espazio honetan dimentsio independente bat da, eta normalean erabiltzen diren sareak milaka edo milioika parametro dituzte [3]. Hortaz, benetan gauzatzen den gradientearen beherapena grafikoki gizakiontzat ulergarria den era batean adieraztea ezinezkoa da.



Irudia 2.6: SGD algoritmoaren adierazpen grafikoa dimentsio bakarreko kasuan (parametro bakarra duen neurona-sarea).

2.3.2 Konboluzio neurona-sareak

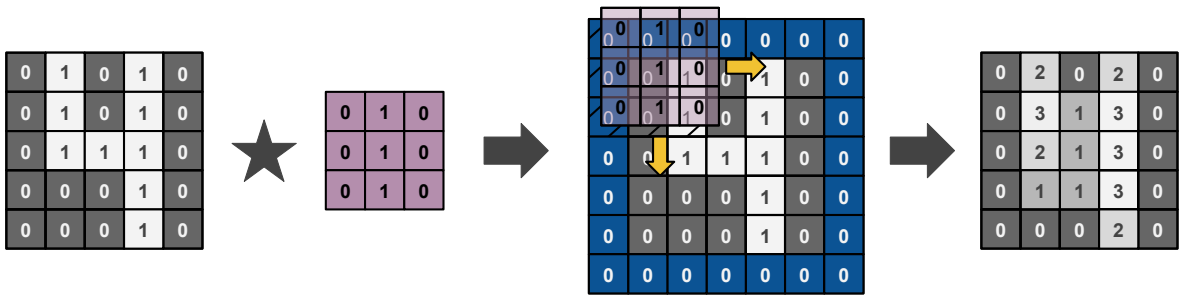
Ikusmen artifizialaren arloan erabiltzen diren neurona-sare askoren oinarrizko blokea konboluzio-eragiketa aplikatzen duen geruza da [24]. Konboluzio neurona-sareak (*Convolutional Neural Network*, CNN) oso ezagunak diren neurona-sare mota bat dira, non konboluzio-geruzek irudietan aplikatuko diren iragazki jakin batzuen multzoa adierazten duten. Konboluzio neurona-sareak zeregin askotan arrakastaz erabiliak izan dira urteetan zehar, testuekin edo

audioekin lan egiteko adibidez, baina batez ere oso hedatuak dira irudi-prozesamenduaren eta ikusmen artifizialaren esparruan [10]. Atal honetan irudiekin lan egiteko erabiltzen diren konboluzio neurona-sareak azalduko dira bakarrik, hauek baitira ikasketa sakonean oinarritutako ikusmen artifizialeko sistema baten funtsezko osagaiak. Hala eta guztiz ere, beste datu mota desberdinekin lan egiteko erabiltzen diren konboluzio neurona-sareak nahiko berdintsuak dira barne-funtzionamenduari dagokionez.

Konboluzio eragiketa Konboluzio-geruzen oinarria den eragiketa hurrengoa da: dimentsio bereko bi tentsoreen elementu-mailako biderkaduraren emaitzaren batuketa [24]. Kanal bakarreko irudiaren kasuan, sarrerako irudia $\mathbf{I}$ eta *kernel* izenaz ezagutzen den iragazkia $\mathbf{K}$ ($m \times n$ dimentsiokoa) izanda, konboluzio-eragiketaren emaitza den *ezaugarri-mapa* $\mathbf{S}$ izango da.

$$\mathbf{S}(i,j) = (\mathbf{I} \star \mathbf{K})(i,j) = \sum_m \sum_n \mathbf{I}(i+m, j+n)\mathbf{K}(m,n) \tag{2.2}$$

Kernel iragazkia sarrerako irudian zehar ezkerretik-eskuinera eta goitik-behera mugitzen da, *stride* izeneko magnitudeak definitzen dituen pausoak eginez, konboluzio-eragiketa sarrerako irudiaren (x, y) koordenatu zehatzetan aplikatuz. Konboluzio-eragiketak ezaugarri-mapa izeneko irudi berri bat sortzen du, non pixel bakoitza jatorrizko irudian posizio jakin batean iragazkia aplikatzearen emaitza den. *Kernel* iragazkia aplikatzeko, iragazkiaren elementu bakoitza sarrerako irudiaren zatiaren pixel baliokidearekin biderkatzen da eta lortutako emaitza guztiak batzen dira. Beraz, konboluzio-geruzen irteerako ezaugarri-mapan pixelak jatorrizko irudiaren pixel originalaren eta honen aldameneko pixelen konbinazio lineal bat izango dira (2.7 Irudia). Sarrerako irudiaren pixel guztietan konboluzio-eragiketa aplikatu ahal izateko, normalean *padding* metodoa erabiltzen da: pixel guztietan *kernel* iragazkia aplikatzea posiblea izan dadin, irudiaren ertzetan 0 balioak gehitzen dira. Horrela, irteerako ezaugarri-mapak sarrerako irudiaren dimentsio espazial berdinak izatea lortzen da [3].



Irudia 2.7: 5×5-ko kanal bakarreko irudi batean ertz bertikalak detektatzen dituen 3×3-ko *kernel* iragazkiaren konboluzio-eragiketaren aplikazioa ($stride = 1$), *padding* erabiliz.

Konboluzio neurona-sareetan, konboluzio-geruza bakoitzak iragazkien multzo bat adierazten du: geruzaren parametroak hainbat *kernel* iragazkien balioak zehazten dituzte. Hori dela eta, konboluzio-geruzek burutu behar duten zeregina ebazteko egokienak diren iragazkiak aurkitzea dute helburu moduan, eta *kernel* iragazki hauek automatikoki bilatzen dituzte entrenamendu-zikloan. Ikasitako iragazki hauek sarrerako irudien ezaugarri zehatzak ateratzeko balio dute. *Kernel* iragazkiak tamaina txikikoak eta karratuak izan ohi dira (3×3 eta 5×5 dimentsio erabilienak dira) eta sarrerako irudiaren sakontasuna dute [24]. Hortaz, sarrerako irudia kolorezko RGB irudia bada, sarearen lehenengo konboluzio-geruzaren 3×3-ko *kernel* iragazkiak ($3 \times 3 \times 3$) dimentsioko tentsoreak izango dira, eta grisen eskalan dagoen irudia bada ($3 \times 3 \times 1$). Kanal bakarreko irudiaren kasuan definitu den konboluzio-eragiketaren adierazpen matematikoa (2.2 Ekuazioa) sakontasun handiagoak dituzten sarreren kasuetarako orokortu daiteke.

$$\mathbf{S}(i,j) = (\mathbf{I} \star \mathbf{K})(i,j) = \sum_{l} \sum_{m} \sum_{n} \mathbf{I}(i+m, j+n, l)\mathbf{K}(m,n,l) \tag{2.3}$$

Konboluzio-geruzen irteeren dimentsio espazialak sarrerarenak izango dira *padding* erabiltzen bada, eta sakontasuna geruza horrek aplikatzen dituen *kernel* iragazkien kopurua izango da [3]. Hori dela eta, neurona-sarearen barnean dauden konboluzio-geruzen iragazkien sakontasuna aurreko geruzak aplikatu dituen *kernel* iragazkien kopurua izango da [24]. Nahiz eta *kernel* iragazkien eta konboluzio-geruzen sarrerako irudien sakontasuna sarean zehar aldatuz joan, iragazki bakoitzarekin lotutako konboluzio-eragiketa aplikatzerakoan lortzen den ezaugarri-mapak kanal bakarra izango du beti. Hurrengo geruzaren sarrera izango den irudia aurreko geruzak sortu dituen ezaugarri-mapak sakontasun-ardatzean bata bestearen ondoren jarriz lortzen da.

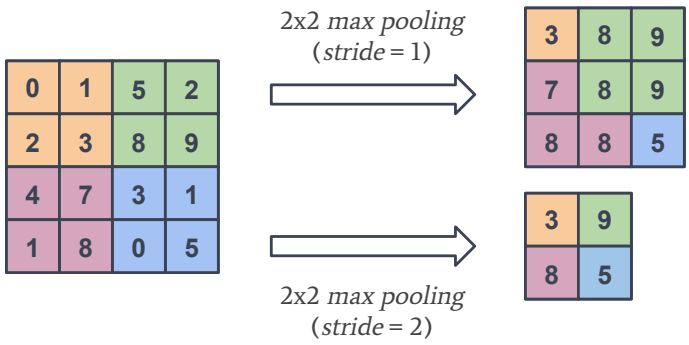
Ikasketa sakonaren hasierako urteetan, dentsoki konektatutako geruzak erabiltzea zen ohikoena. Geruza hauen eta konboluzio-geruzen arteko desberdintasun nagusia hurrengoa da: dentsoki konektatutako geruzek sarrerako datuen ezaugarri globalak ikasten dituzte; konboluzio-geruzek, berriz, ezaugarri lokalak ikasteko erraztasun handiagoa dute [3]. Arrazoi honengatik, konboluzio neurona-sareek bi propietate oso interesgarri dituzte:

- Ikasten dituzten barne-errepresentazioak translazioekiko aldaezinak dira, eta gure errealitatearen mundu bisuala funtsean translazioekiko aldaezina da ere [3]. Irudi baten zonalde jakin batean agertzen den ezaugarri bisual bat ikasi ezkero, ezaugarri hori edonon igartzeko ahalmena garatzen dute. Hau da, ezaugarrien lokaltasuna kontuan izateko gai dira [10]. Dentsoki konektatutako geruza batek, ordea, lehen ikasitako ezaugarri berdina posizio berri batean aurkitu ezkero, ezaugarria berriro hasieratik ikasi beharko luke. Hori dela eta, konboluzio neurona-sareak oso eraginkorrak dira irudiak prozesatu behar direnean, eta orokortze-ahalmen handia duten errepresentazioak ikasteko entrenamendu-zikloan laginen kopuru txikiagoa eta denbora gutxiago behar dute beste sare mota batzuekin konparatuz [3].

- Ezaugarrien hierarkia espaziala ikasi ahal dute, eta mundu bisuala espazialki hierarkikoa da. Sarearen hasieran dagoen konboluzio-geruzak tamaina txikiko ezaugarri lokalak igartzen ikasiko du, ertzak esaterako. Sarearen amaieran dagoen geruzak, ordea, abstraktuagoak diren ezaugarri handiagoak ikasiko ditu, hasierako konboluzio-geruzek ikasitako ezaugarri txikiagoez osatuak [24]. Orokorrean, sarearen lehenengo geruzak sarrerako irudiaren ertz mota jakin batzuk (horizontalak, bertikalak edo izkinak) detektatzen dituzten iragazkiak eraikitzen ditu. Bigarren geruza batek lehenengo geruzaren irteeretan oinarrituz forma geometriko sinpleak (kurbak, zirkuluak, laukizuzenak, triangeluak eta antzekoak) antzematen dituzten iragazkiak lortu ahal ditu. Hurrengo geruzek aurreko geruzen irteera sinpleagoen konbinazioetan oinarrituz goi-mailako ezaugarri konplexuagoak (aurpegiak, katuak, autoak etab.) sortzeko ahalmena garatzen dute. Mekanismo honetaz baliatuz, konboluzio neurona-sareek era eraginkorrean gero eta konplexuagoak eta abstraktuagoak diren kontzeptu bisualak ikasi ahal dituzte [3].

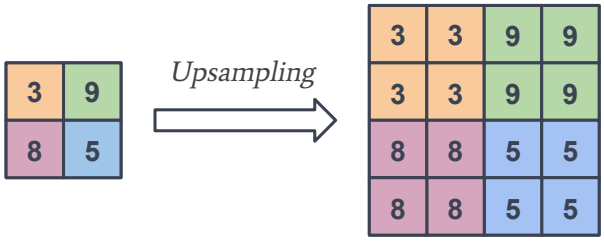
***Pooling* geruzak** Konboluzio neurona-sareen arkitekturetan konboluzio-geruzen artean *pooling* geruzak jartzea ohikoa da. *Pooling* geruzen funtsezko helburua neurona-sarea zeharkatzen duen irudiaren dimentsio espazialak (altuera eta zabalera) mailaka murriztea da, horrela sareak dituen parametroen kopurua, zama konputazionala eta, ondorioz, sarea entrenatzeko beharrezkoa den denbora txikitzeko [24].

Pooling geruzek sarrerako irudiaren sakontasun kanal bakoitzean independenteki aplikatzen dira, eta ohikoenak *max pooling* eta *average pooling* dira. *Max pooling* eragiketak irudian zehar pausoz pauso (*stride*) mugituz sarrerako iruditik $m \times m$ dimentsioko blokeak hartzen ditu (konboluzio-geruzek egiten duten moduan), eta bloke horietan agertzen den balio handienarekin bakarrik geratzen da. *Average pooling* eragiketak, berriz, blokean agertzen diren balioen batezbestekoa kalkulatzen du. Normalean *pooling* geruzek 2×2-ko blokeak (*pool size*) erabiltzen dituzte $stride = 2$ pausoarekin, horrela sarrerako irudiaren dimentsio espazialak erdira txikituz. Lan honetan diseinatu diren neurona-sareetan 2×2-ko *max pooling* geruzak $stride = 2$ pausoarekin erabiliko dira.



Irudia 2.8: 4×4-ko kanal bakarreko irudi batean 2×2-ko *max pooling* eragiketa aplikatzearen emaitza, $stride = 1$ eta $stride = 2$ erabiliz.

***Upsampling* geruzak** Neurona-sare diseinu batzuetan *pooling* geruzak erabiliz txikitu diren irudien dimentsio espazialak sarean aurrera egin ahala berriro berreskuratu nahi dira. Horretarako, *upsampling* geruzen erabilera metodo ohikoena da [11]. Geruza hauek *nearest neighbour* algoritmoaren bidez irudien bereizmena bikoizteko balio dute (2.9 Irudia).



Irudia 2.9: 2×2-ko irudi baten bereizmena *nearest neighbour* algoritmoaren bidez bikoiztu daiteke, *upsampling* geruzetan erabiltzen den metodoa.

2.3.3 Galera-funtzioak

Ikasketa automatikoan, galera-funtzioa ikasketa-prozesua norabide egokian gidatzen duen funtzioa da, hau da, neurona-sarearen parametroak nola doitu behar diren zehazten duen metrika da, sareak burutu behar duen zeregina hobeto ebatzi dezan. Zehazki, galera-funtzioak neurona-sareak sortutako irteera eta itxarondako benetako balioa konparatzen ditu eta diferentzia honetan oinarrituz sarearen parametroak galera-balioa txikitzen duen norabidean doitzen dira, horrela neurona-sarearen irteera eta itxarondako balioa gerturatuz. Beraz, sarearen portaeran oso eragin handia duen diseinu-erabaki garrantzitsua da galera-funtzioaren aukera.

Lehen aipatu den bezala, galera-funtzioaren txikitzea *Stochastic Gradient Descent* (SGD) izenaz ezagutzen den algoritmoaren bidez egiten da normalean, gradientearen beherapenaren kontzeptuan oinarrituta dagoen algoritmoa. Hau da, neurona-sarea osatzen duten geruzen pisuak (sarearen parametroak) hauek sorrarazten duten galera-funtzioaren gradientearen kontrako norabidean doitzen dituen algoritmoa da. Hurrengoak lan honetan zehar erabiliko diren galera-funtzioen definizioak dira.

Mean Squared Error (MSE) *Mean Squared Error* (MSE) erregresio-problemetan gehienetan erabiltzen den galera-funtzioa da. Problema mota hauetan helburua balio erreal bat iragartzea da. Itxarondako benetako balioak banaketa normal batetik hurbil daudenean errendimendu ona erakusten duen galera-funtzioa da [2]. Galera-funtzio honek irteerako tentsorearen eta itxarondako benetako tentsorearen elementuen diferentzien karratuen batezbesteko balioan oinarrituta dago. Beraz, galera-balioa beti da positiboa eta balio perfektua 0 izango da. Diferentzien karratua erabiltzen denez, errore handiek errore txikiek baino galera-balio proportzionalki handiagoak sortuko dituzte, horrela neurona-sareak normalean itxarondako balioetatik gehiegi urruntzen diren irteerak bueltatzea gehiago zigortuz.

$$\mathcal{L}_{MSE}(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2 \tag{2.4}$$

Mean Absolute Error (MAE) Erregresio-problema batzuetan, itxarondako balio gehienak banaketa normal baten barnean egon arren, iragarri nahi diren hainbat balio atipikoak dira. Kasu honetan MSE galera-funtzioa aplikatzen bada, itxarondako benetako balioetan agertzen diren balio atipiko horiek iragartzeko gaitasuna asko ahuldu daiteke. Hortaz, balio atipikoak garrantzitsuak eta adierazgarriak diren problemetan, balio hauek hobeto mantenduko dituen *Mean Absolute Error* (MAE) galera-funtzioa erabiltzea ohikoa da [2]. Funtzio hau irteerako tentsorearen eta itxarondako benetako tentsorearen elementuen diferentzien balio absolutuaren batezbesteko balioan oinarrituta dago. MSE funtzioarekin gertatzen den bezala, 0 baliora iristeak neurona-sarearen irteera hobeezina dela esan nahi du.

$$\mathcal{L}_{MAE}(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}_i - Y_i| \tag{2.5}$$

Binary Cross-Entropy (BCE) *Cross-Entropy* informazio-teoriaren esparruan maiz erabiltzen den neurketa da: bi probabilitate-banaketen arteko entropia totala kalkulatzen du, edo beste era batean esanda, probabilitate-banaketa hauen arteko diferentzia [2]. *Cross-Entropy* galera-funtzio bezala erabiltzea ohikoa da sailkapen-problemetan. Galera-funtzio honen bidez bi tentsore konparatzeko gai izateko, tentsore hauek bi probabilitate-banaketa desberdin adierazten dituztela onartu behar da. *Binary Cross-Entropy* galera-funtzioa klase bakarreko sailkapen-problemetan erabili ohi da eta klase bat bakarrik duten segmentazio-mapak irteera bezala bueltatzen dituzten neurona-sareekin lan egitean erabili daiteke (ikusi 4.1 Atala). Mota honetako mapen pixelen intentsitateak 0 eta 1 balioen artean normalizatu daitezke, horrela probabilitate bezala interpretatu ahal izateko. Beraz, kasu honetan irteerako tentsorearen elementuen balioak $\widehat{Y}_i \in [0, 1]$ balioen artean egongo dira, eta benetan itxarondako tentsorearen elementuen balioak $Y_i \in \{0, 1\}$ (bakarrik bi balio posible, klase bakarra) modukoak izango dira. Berriro ere, galera-funtzio hau minimizatu beharreko funtzioa da eta galera-balio ideala 0 izango da.

$$\mathcal{L}_{BCE}(Y, \widehat{Y}) = -\sum_{i=1}^{n} Y_i log(\widehat{Y}_i) + (1 - Y_i) log(1 - \widehat{Y}_i) \tag{2.6}$$

3 Ikusmen artifiziala

Ikusmen artifizialaren arloa 1970ko hamarkadan jaio zen, agente adimentsu orokorren sorkuntza helburu bezala zuen ahalegin kolektibo baten barnean. Hasiera batean, adimen artifizialaren eremuko aurrendari batzuk robotek giza-adimena eskuratzeko ebatzi behar ziren problemen artean ikusmen-pertzepzioaren zeregina erlatiboki erraza izango zela uste zuten. Gaur egun, zeregin honen konplexutasun-maila garai hartan pentsatzen zena baino askoz handiagoa dela dakigu [27].

Ordenagailu bidezko ikusmen artifizialaren esparruaren funtsezko helburua ordenagailuek datu bisualetatik abiatuz goi-mailako ulermena garatzea da. Hau lortzeko, eszena bisual batean agertzen diren elementuak detektatu, identifikatu, sailkatu eta antolatu behar dira, prozesu honetan jakintza abstraktua eta hainbat zeregin desberdin (aurpegi-sailkapena, irudien zaharberritzea edo segmentazio semantikoa, esaterako) burutzeko ahalmena irabaziz [24]. Hori dela eta, gizakiok inguruko munduaren 3 dimentsioko egitura arazorik gabe automatikoki hautemateko dugun gaitasuna ikusmen artifizialeko sistemetan erabili ahal izatea oso onuragarria izango litzateke zeregin mota asko hobeto burutu ahal izateko. Hala ere, giza-pertzepzioaren atzean dagoen ikusmen-sistemak lan egiteko duen moduaren azalpena pertzepzioaren arloko psikologoek gaur egun oraindik guztiz ebatzi ez duten problema da [27].

Beraz, arlo honetan helburu garrantzitsuena irudietatik informazio erabilgarria ateratzea da, algoritmoak bete behar duen zeregina burutzea errazten duten ezaugarri abstraktu gisa (ertzak, koloreak, testurak etab.). Tradizionalki, beharrezkoak diren irudien ezaugarri garrantzitsu hauek eskuz definitu izan dira eta, hainbat algoritmo desberdinen bidez, ezaugarri zehatz hauek erabiliz problema mota ezberdinak ebazteko egokiak diren ikusmen-sistemak diseinatu dira.

Halaber, azkenengo urte hauetan ikusmen artifizialeko arloan aurrerapen nabarmenak egin dira, batez ere ikasketa sakonaren bidez lortuak. Ikasketa sakoneko arkitekturak irudietatik informazio erabilgarria lortzeko beharrezkoak diren ezaugarri abstraktuen ateratzea modu automatiko batean egiteko gai dira eta, gainera, aldi berean ezaugarri hauek lortzeko eta zeregina ebazten ikasteko ahalmena dute [6]. Neurona-sareek beste abantailak dituzte ere: ondoz ondoko geruzez osatutako arkitektura sakonak direnez, beste ikasketa automatikoko algoritmoekin alderatuz sarrerako irudietatik abiatuz barne-errepresentazio abstraktuagoak eta baliagarriagoak eratzeko gaitasuna dute, adibidez [3]. Hau guztia dela eta, 2.3.2 Atalean aztertu diren konboluzio neurona-sareak ikusmen artifizialeko sistemak eraikitzerako orduan funtsezko osagaietan bihurtu dira.

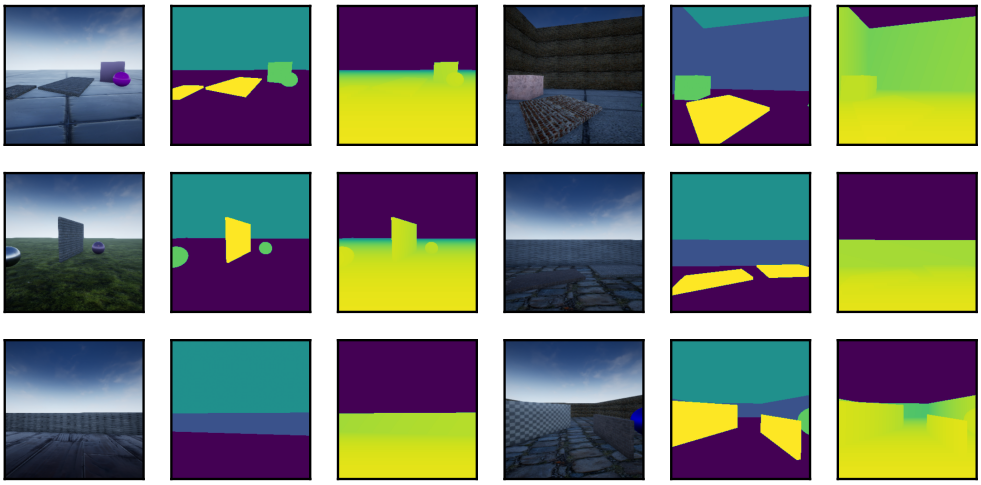
3.1 Irudien tentsore-adierazpena

Irudiekin lan egiten duen sistema bat eraikitzerako orduan, lehenengo pausoa irudi bat matematikoki nola adierazten den ulertzea da. Horretarako, irudien oinarrizko blokeak diren pixelekin hasi behar da. Normalean pixelek irudi baten posizio zehatz batean agertzen den

argiaren kolorea eta intentsitatea adierazten dute. Irudi digital guztiak pixelez osatuta daude: irudia lauki-sare bat bezala adierazi ezkero, sarearen lauki bakoitza pixel bat izango litzateke [24].

Pixel gehienak bi modu desberdinetan adierazi ohi dira: grisen eskalan edo kolorez. Zuribeltzean dauden irudiak grisen eskalan daudela esaten da, irudia tentsore bat bezala adierazten badugu (`altuera` $\times$ `zabalera` $\times$ `1`) forma izango du, sakontasun-kanal bakarra du. Kolorezko irudiak, berriz, RGB kolore-espazioan adierazi ohi dira eta tentsore moduan (`altuera` $\times$ `zabalera` $\times$ `3`) forma dute, sakontasun-kanal bat kolore (gorria, berdea, urdina) bakoitzeko. Beraz, grisen eskalan dauden irudien pixelak $[0, 255]$ (0 beltza eta 255 zuria) artean dagoen zenbaki arrunt bakarra izango dute eta kolorezko pixelak, ordea, mota horretako 3 balio dituzten bektoreak izango dira. Hortaz, `(red, green, blue)` bektoreak RGB kolore-espazioko kolore jakin bat adieraziko du.

Irudi batekin tentsore bezala lan egiteko, `(altuera, zabalera, sakontasuna)` formako *NumPy array* bat erabili ohi da *Python*-en. Altuera zabalera baino lehenago ezartzen da matrizeetan elementuak eskuratzeko (`lerroak` $\times$ `zutabeak`) notazioarekin bat etortzeko [24], nahiz eta normalean irudiez hitz egiterakoan kontrako ordena erabiltzen den. Izan ere, irudi baten adierazpen tentsorialaren lerroen kopurua bere altuera zehazten du eta zutabeena, berriz, bere zabalera.



Irudia 3.1: *IntPhys* datu-multzoan eskuragai dauden laginen adibideak (kolorezko RGB irudiak, segmentazio-mapak eta sakontasun-mapak).

3.2 *IntPhys*

Adimen artifizialeko sistemek zeregin bisual konplexuetan giza-errendimendua eskuratzeko, lehenengo gure errealitate fisikoan agertzen diren objektu makroskopikoak, mugimenduak eta indarrak ulertzeko gaitasuna lortu behar dute. *IntPhys*[1] [22] ikusmen artifizialeko sistemek oinarrizko intuizio fisikoa garatzeko helburuarekin diseinatua dagoen datu-multzoa da, umeek oso txikiak direnean mundu makroskopikoko fisika newtondarra barneratzeko duten eran inspiratua. Horrela, datu-multzo hau hainbat objektu makroskopikoen zinematika eta dinamika erakusten dituzten eszena bisualen bilduma bezala deskribatu daiteke. Beraz, umeen moduan esperientzien errepikapenaren ondorioz ere ikasten duten adimen artifizialeko sistemak entrenatzeko ezinbestekoa den laginen kopuru handia eskaintzen du *IntPhys*-ek.

IntPhys datu-multzoa *Unreal Engine*[2] bideo-jokoak sortzeko tresna erabiliz sortu diren bideo sintetikoz osatua dago. Bideo bakoitzak 100 fotograma ditu eta hainbat mota desberdineko objektuen mugimenduak eta hauek ingurunearekin edo haien artean dituzten elkarrekintza fisikoak adierazten dituzte. Bestalde, bideoak 15 $fotograma/segundo$ formatuan simulatuta daude eta, ondorioz, eszena bakoitzaren luzera $\sim$ 7 segundokoa da. Gainera, bideoetan agertzen diren objektuak gure errealitatearen fisikaren legeak jarraitzeko programatuta daude, *Unreal Engine* barnean dagoen *PhysX*[3] simulagailu fisikoa erabiliz.

Eszena bisual bakoitzean $\{1, 2, 3\}$ objektu eta $\{0, 1, 2\}$ estaldura dinamiko agertu ahal dira. Objektuak estatikoak ala dinamikoak izan daitezke eta hainbat gainazaleko testurak (plastikoa, metala, egurra etab.), koloreak (kolore biziak ala itzaliak) eta formak (esferak, kuboak, konoak etab.) dituzte. Estaldura dinamikoak beti daude mugimenduan, eta haien helburu nagusia objektuak ezkutatuz neurona-sareen iragartze-zeregina zailtzea da. Haien artean hain desberdinak izan daitezkeen objektuak erabiltzean, sareek objektuen kontzeptu orokorragoa ikastea nahi da. Adibidez, askoz errazagoa da berde bizi bateko gainazala duen esfera objektu bezala identifikatzea ingurunea islatzen duen metalezko esfera bat baino.

Neurona-sareak entrenatzeko 15.000 bideo eskuragai daude, eta bideo bakoitza 100 fotogramaz osatuta dagoenez datu-multzoak guztira 1.500.000 irudi ditu. Bideo bakoitza kolorezko RGB irudien (288×288 pixel) sekuentzia bezala adierazten da, guztira 157Gb betez. Kolorezko irudi bakoitzarekin lotuak, neurona-sareen entrenamendua errazteko beste bi irudi eskaintzen dira ere: segmentazio-mapa eta sakontasun-mapa (3.1 Irudia). Segmentazio-mapak fotograma osatzen duten osagai kontzeptualak banantzen ditu: zorua, paretak, zerua, objektuak eta estaldurak. Sakontasun-mapa, berriz, pixel bakoitzaren posiziotik kameraraino dagoen distantzia kodifikatzen du. Bestalde, lehen aipatu den bezala bideo guzti hauetan agertzen diren objektuen mugimenduak eta haien dinamikak guztiz fisikoki posibleak dira.

[1] IntPhys: https://www.intphys.com
[2] Unreal Engine: https://www.unrealengine.com
[3] PhysX: https://developer.nvidia.com/physx-sdk

4 Segmentazio semantikoa

Konboluzio neurona-sareen erabilera hedatuena irudi-sailkapena egitea da eta sare mota hauen irteera etiketa bakarra izan ohi da, normalean zenbaki arrunt bat. Hala ere, ikusmen artifizialeko zeregin jakin batzuetan sarearen irteera irudi oso bat izatea egokiagoa izan daiteke, eta horretarako sarrerako irudiaren pixel-mailako sailkapena egitea beharrezkoa izaten da. Beraz, irudien segmentazio semantikoaren zereginean sarearen irteeran sarrerako irudiaren pixel bakoitzari etiketa bat ezartzen zaio, hau da, pixel bakoitza klase zehatz batean sailkatzen da. Segmentazio semantikoan sarrerako irudiaren pixel bakoitzarentzat iragarpen bat egiten denez, zeregin hau iragarpen dentso izenarekin ezagutzen da ere. Teknika hau biomedikuntzaren arloan maiz erabiltzen da eta esparru horren irudi-prozesamenduan bereziki garrantzi handia dauka [23]. Irteera bezala irudi oso bat behar duten beste zeregin batzuen adibideak ondorengo atalean landuko den etorkizuneko fotogramaren iragarpena eta sakontasunaren balioespena dira.

Lan honetan inplementatzen den segmentazio semantikoan klaseen instantziak ez dira bereizten; pixel bakoitzaren sailkapena egiten da soilik. Hori dela eta, sarrerako irudi batean objektu-klaseko bi objektu agertzen badira, segmentazio-mapak ez ditu objektu horiek bi elementu desberdin bezala identifikatuko, objektu-klasearen barnean dauden pixel guztiak adieraziko ditu bakarrik [11]. Hala ere, instantzia-segmentazioa egiten dituzten beste teknika batzuk existitzen dira, baldintza hau burutu nahi den zereginean ezinbestekoa izan ezkero, baina inplementazio zailagoa dute orokorrean eta sarea entrenatzeko instantzia-segmentaziorako bereziki prestatutako datu-multzo zehatzak behar dituzte [9].

4.1 Helburua

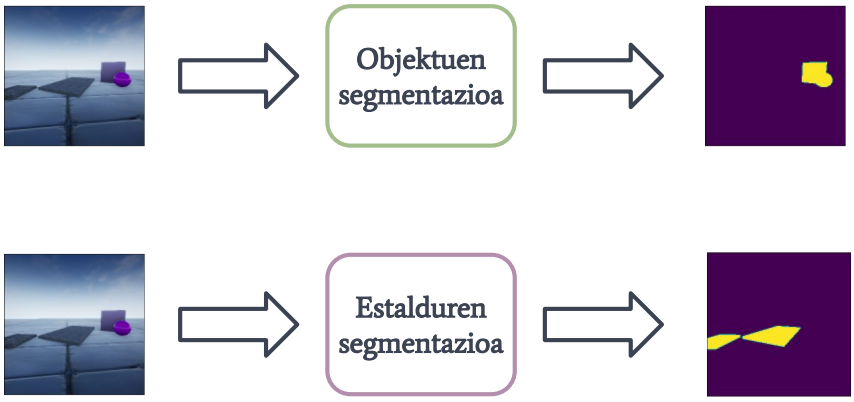
Segmentazio semantikoaren helburua hurrengoa da: sarrera bezala kolorezko RGB irudia (`altuera` × `zabalera` × `3`) edo grisen eskalan dagoen irudia (`altuera` × `zabalera` × `1`) hartuz, irteeran sarrerako irudiaren pixel bakoitza klase jakin batean sailkatzen duen *segmentazio-mapa* bat lortzea. Hortaz, segmentazio semantikoaren zereginean ondorengo galderaren erantzuna aurkitu nahi da: kontzeptualki zer agertzen da irudi batean eta non dago hori espazialki kokatua?

Sailkapen-zereginetan ohikoa den moduan, neurona-sarearen itxarondako irteeran agertu ahal diren klaseen etiketak *one-hot* eran kodifikatuz definitzen dira. Beste era batean esanda, klase posible bakoitzari irteerako segmentazio-maparen sakontasun-kanal bat dagokio: kanal bakoitzean sarrerako irudian klase horretakoak diren pixel guztiak 1 zenbakiaz adierazten dira, gainerako beste pixel guztiak 0 izanda [11].

Horrela, sarearen irteera bezala lortzen den segmentazio-maparen pixel bakoitzean, balio handiena duen kanalaren posizioa pixel zehatz horren klasea zehaztuko du. Beste era batean esanda, irteerako segmentazio-maparen pixel bakoitzari dagokion sakontasun-bektorearen balio maximoaren posizioa kalkulatuz pixel horren klasea zehaztu daiteke. Hori dela eta,

segmentazio semantiko mota honetan pixel guztiek klase bakarra izan dezakete gehienez. Bestalde, neurona-sarearen irteerako segmentazio-maparen sakontasun-kanal bat aukeratzean, sarrerako irudian kanal horrekin lotuta dagoen klaseari dagokion azalera adierazten duen *segmentazio-maskara* lortzen da [11].

Lan honetan segmentazio semantikoa burutzen duten bi neurona-sare entrenatuko dira: objektuak segmentatzen dituen sarea eta estaldurak segmentatzen dituena. Bi hauek *IntPhys* [22] datu-multzoaren bideoetan agertzen diren elementu dinamikoak dira, eta haien segmentazio-maskarak hurrengo atalean landuko den etorkizuneko fotogramaren iragarpenerako erabilgarriak izan daitezke. Lan honetan segmentazio semantikoa burutzen duten bi neurona-sare entrenatzea erabaki da, sare bakoitza elementu bisual zehatz batean hobeto espezializatu ahal izateko. Horren ondorioz, neurona-sare hauen irteerak kanal bakarreko segmentazio-mapak izango dira, bakoitzak klase bakarra sailkatzen du eta.



Irudia 4.1: Objektuen eta estalduren segmentazio-mapak entrenatu diren bi neurona-sareen helburuak dira.

4.2 Sarearen arkitektura

Irudien segmentazio semantikoa egiteko neurona-sare bat diseinatzerako orduan, burura etorri ahal den lehenengo arkitektura sinplea sarrerako irudiaren dimentsio espazialak mantentzen dituzten konboluzio-geruza batzuk bata bestearen atzean jartzean lortzen den neurona-sarea izan daiteke. Izan ere, soilik konboluzio-geruzez osatutako neurona-sareek irudiekin lan egiterako orduan orokorrean dentsoki konektatutako geruzak erabiltzen dituzten sareak baino emaitza hobeagoak lortzen dituzte [26]. Hala ere, sare osoan zehar sarrerako irudiaren jatorrizko tamaina eta bereizmena mantentzeko parametroen kopuru handia beharrezkoa da, eta hau konputazionalki garestia izan ahal da konboluzio-geruzen kopurua oso txikia ez bada [11].

Honen ondorioz, diseinu mota hau erabiltzen dituzten arkitekturek ezin dira oso sakonak izan eta arrazoi honengatik askotan irudien ezaugarri abstraktuak lortzeko arazoak izaten dituzte.

Izan ere, 2.3.2 Atalean aipatu den bezala konboluzio neurona-sareetan hasierako geruzek behemailako kontzeptuak adierazten dituzten ezaugarriak ikasten dituzte eta sarearen bukaeran dauden geruzek goi-mailako ezaugarri abstraktuagoak ateratzen dituzte, honetarako aurreko konboluzio-geruzek lortu dituzten ezaugarri sinpleagoak konbinatuz. Gainera, azkenengo geruzetan ateratzen diren ezaugarri konplexuagoen adierazkortasuna hobetzeko, sarean zehar dauden konboluzio-geruzen iragazkien kopurua gradualki handitzea beharrezkoa da [11].

Irudien sailkapenaren zereginean, aurreko mugek ez dute arazo handirik suposatzen: kasu horretan irudiaren eduki kontzeptuala, baina ez espaziala, interesa du bakarrik [11]. Beraz, sarrerako irudiaren dimentsio espazialak (altuera eta zabalera) txikitzen dituzten *pooling* geruzak sarean periodikoki erabiltzearekin nahikoa da parametroen kopurua kontrolatzeko eta zama konputazionala murrizteko. Beste era batean esanda, datuen bereizmen espaziala murrizten da parametroen kopuru bera erabiliz neurona-sarearen geruzen kopurua eta sakontasuna handitu ahal izateko. Irudi-sailkapena ez diren beste zeregin batzuetan, berriz, sarrerako irudiaren eduki kontzeptuala *eta* espaziala interesa dute; irudien segmentazio semantikoan edo sakontasunaren balioespenean, esaterako. Kasu hauetan sarearen irteera sarrerako irudiaren dimentsio berdinak dituen beste irudi bat izango da, eta neurona-sarearen barnean irudi-datuen dimentsio espaziala berreskuratzeko metodoren bat beharrezkoa da.

***U-Net* arkitektura** Azkenengo urte hauetan irudien segmentazio semantikoaren arloan arrakasta handia izan duen arkitektura *U-Net* [23] izenaz ezagutzen den konboluzio neurona-sarea da. Neurona-sare hau biomedikuntzan mikroskopio baten bidez lortzen diren zelulen irudiak automatikoki segmentatzeko helburuarekin sortu zen, baina horrez gain esparru horretatik kanpo oso arrakastatsu bihurtu zen berehala. Izan ere, sare honek oso emaitza onak lortzen ditu, eredua entrenatzeko eskuragai dauden datuen kopurua txikia denean ere, eta azkarra izateaz gain ez da inplementatzeko konplexuegia.

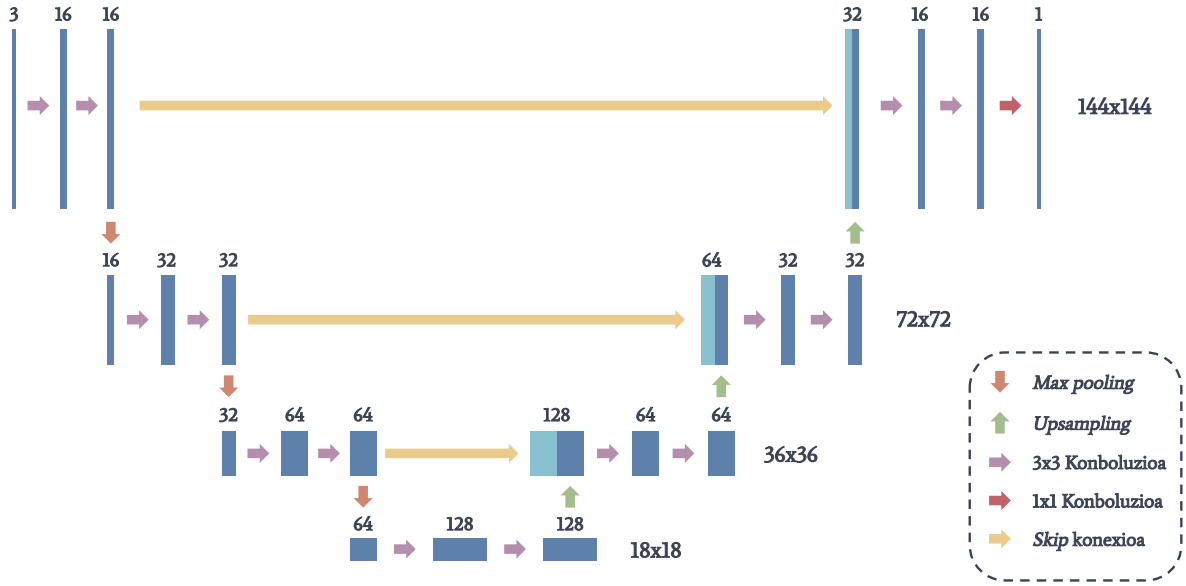
U-Net arkitektura *Fully Convolutional Neural Network* (FCNN) [26] izenaz ezagutzen den arkitekturan oinarrituta dago. Sare-diseinu honi egindako aldaketak eta gehiketak *U-Net* arkitekturan oinarritutako neurona-sareek irudi-laginen kopuru txikiagoekin segmentazio zehatzagoak lortzea posible egiten dute [23]. FCNN sarearen atzean dagoen funtsezko ideia hurrengoa da: normalean irudien sailkapenean ohikoa den konboluzio eta *pooling* geruzez egindako *kodetzailea* irteerako irudiaren dimentsio espazialak berriro berreskuratzen dituen *deskodetzaile* batekin osatzea, non *pooling* geruzak *upsampling* geruzekin ordezkatzen diren. Beraz, geruza berri hauek sarea zeharkatzen duten irudien bereizmena handitzen dute berriro. Lokalizazioa hobetzeko, kodetzailearen konboluzio-geruzek lortzen dituzten ezaugarriak deskodetzailearen *upsampling* geruzen irteerekin konbinatzen dira *skip* konexioak erabiliz, honetarako bien arteko kateamendu bat eginez. Beraz, deskodetzailearen konboluzio-geruzek irteera egokiagoak ikasi ditzakete bi informazio-iturri hauek bateratzearen ondorioz.

U-Net arkitekturan, sarearen kodetzailea eta deskodetzailea simetrikoak dira eta, beraz, deskodetzailearen konboluzio-geruzek ere iragazkien kopuru handia dute, kodetzailearen geruzetan gertatzen den bezala. Honen ondorioz, U-formako sare honek testuinguru-informazioa eta ezaugarrien lokaltasuna bereizmen handiagoko hasierako geruzetatik neurona-sarearen amaieran

dauden geruzetara erraztasunez hedatzeko gaitasuna dauka, *skip* konexioei esker [23]. Gainera, arkitektura honek ez ditu dentsoki konektatutako geruzarik erabiltzen, irudiekin lan egiterako orduan hauek behar dituzten parametroen kopurua handiegia delako, konboluzio-geruzekin alderatu ezkero [26].

***U-Net-Seg* arkitektura** Lan honetan *U-Net* sarean oinarritutako sare bat diseinatu da segmentazio zereginak burutzeko: 4.2 Irudian adierazita dagoen *U-Net-Seg* izeneko sarea. Kodetzaileak irudien sailkapenerako erabiltzen diren konboluzio neurona-sareek jarraitu ohi duten egitura dauka: Bi 3×3-ko konboluzio-geruzez eta *stride* $= 2$ duen 2×2-ko *max pooling* geruza batez osatutako blokea 3 aldiz aplikatzen da. *Max pooling* geruzen bidez, irudiaren bereizmena erdira murrizten da bloke bakoitzaren irteeran. Gainera, blokeek osatzen dituzten konboluzio-geruzek aurreko blokearen geruzek dituzten iragazkien kopurua bikoizten dute aplikazio bakoitzean.

Bestalde, deskodetzailearen blokeak irudiaren bereizmena bikoizten dituzten 2×2-ko *upsampling* geruza batekin hasten dira, ondoren iragazkiak erdira murrizten dituen 2×2-ko konboluzio-geruza aplikatzen da eta kodetzailearen bloke baliokidearen azkenengo konboluzio-geruzaren irteera kateamendu baten bidez lotzen da (*skip* konexioak). Gero, aurreko blokearen geruzek duten iragazkien kopuru erdia dituzten bi 3×3-ko konboluzio-geruzak aplikatzen dira. Bukatzeko, klase bakarreko segmentazio-mapa lortzeko iragazki bat duen 1×1-ko konboluzio-geruza erabiltzen da.



Irudia 4.2: Irudien segmentazio semantikorako erabiltzen den *U-Net-Seg* sarearen arkitekturaren eskema orokorra.

Dice koefizientea Irudien segmentazio semantikoaren arloan arrakasta handia duen galera-funtzioa *Dice* koefizientean oinarritua dago [13]. Datu boolearrekin ($\{0, 1\}$) lan egitean ohikoak diren benetako positiboen (*True Positive*, TP), positibo faltsuen (*False Positive*, FP) eta negatibo faltsuen (*False Negative*, FN) definizioak erabiliz, *Dice* koefizientea hurrengo moduan definitu daiteke:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{4.1}$$

Dice koefizientearen bidez, bi segmentazio-mapek duten gainezartze-maila neurtu daiteke: 0 balioak neurona-sareak egindako iragarpenaren pixel guztien sailkapena gaizki dagoela esan nahi du, 1 balioa lortzean, ordea, bi segmentazio-mapak guztiz identikoak dira (iragarpenean negatibo faltsuak edo positibo faltsuak ez dira agertzen, pixel-sailkapen guztiak benetako positiboak eta benetako negatiboak dira). Beraz, *Dice* koefizientea $[0, 1]$ arteko balio bat bueltatzen du. Minimizatu behar den galera-funtzio batean bihurtzeko ondorengoa egin ohi da:

$$\mathcal{L}_{DICE} = 1 - \frac{2TP}{2TP + FP + FN} \tag{4.2}$$

Normalean, pixel-mailako beste galera-funtzio batekin konbinatzen da, *Dice* koefizienteak segmentazio-maparen egoera globala kontuan hartzen duelako bakarrik [13]. Kasu honetan klase bakarreko segmentazio semantikoa burutzen denez, *Binary Cross-Entropy* galera-funtzioarekin konbinatuko da *U-Net-Seg* neurona-sarea segmentazio-zereginean entrenatzeko erabiliko den galera-funtzioa lortzeko.

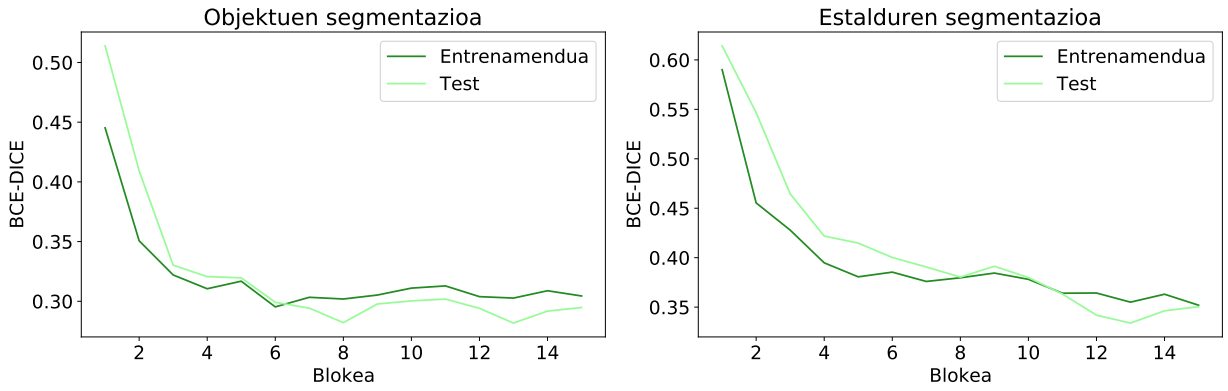
$$\mathcal{L}_{BCE-DICE} = \lambda_{BCE}\mathcal{L}_{BCE} + \lambda_{DICE}\mathcal{L}_{DICE} \tag{4.3}$$

λ_{BCE} eta λ_{DICE} parametroak doituz galera-funtzio bakoitzak amaierako galera-balioan izango duen eragina kontrolatu daiteke. Lan honen kasuan, $\lambda_{BCE} = 0.5$ eta $\lambda_{DICE} = 0.5$ balio orekatuak aukeratzean emaitza onak lortzen direla behatu da.

4.3 Entrenamendua

IntPhys datu-multzoan eskuragai dauden 15.000 bideoak 1.000 bideoko 15 bloke desberdinetan banandu dira, *Google Colab*[4]-en mugak direla eta 157Gb betetzen dituen datu-multzo osoa aldi berean erabiltzea ezinezkoa izan delako. Gainera, bloke bakoitzaren bideoak beste bi multzoetan banandu dira ere: 900 bideo entrenamendu-zikloan erabiliko dira sarea entrenatzeko (entrenamendu-multzoa) eta geratzen diren 100 bideoak neurona-sareak lortu duen orokortze-ahalmena neurtzeko erabiliko dira (test-multzoa).

Neurona-sarea entrenatzeko *Adam* optimizatzailea [12] erabili da, $\epsilon = 0.001$-ko ikasketa-erritmoarekin. Bestalde, entrenamendu-zikloaren luzera murrizteko helburuarekin sarrerako irudien bereizmena erdira jaitsi da, (288×288)-tik (144×144)-ra pasatuz. *U-Net-Seg* neurona-sarea bloke bakoitzean behin bakarrik entrenatzen da, eta sareak bloke baten entrenamendu-zikloa amaitzean lortzen dituen pisuen balioak gorde eta hurrengo blokean entrenamendua hasi baino lehen berriro ezartzen dira. Arkitektura honek 485.297 parametro ditu guztira eta bloke bakoitzean entrenamendu-zikloa burutzeko $\sim$ 21 minutu behar dira (guztira 5 ordu eta 15 minutu behar dira entrenamendu-multzo osoan sarea entrenatzeko). Entrenamendu-zikloan erabili den galera-funtzioa aurreko atalean definitutako $\mathcal{L}_{BCE-DICE}$ funtzioa da eta bi *U-Net-Seg* sareek entrenamenduan zehar eta bloke bakoitzaren test-multzoetan lortu dituzten galera-balioak 4.3 Irudian adierazten dira.



Irudia 4.3: Irudien segmentazio semantikorako erabiltzen diren bi neurona-sareek entrenamenduan zehar eta test-multzoetan izandako galera-funtzioaren balioak.

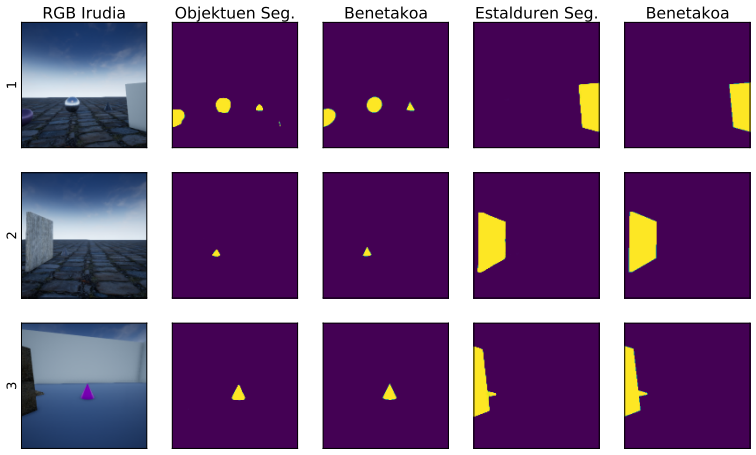
Overfitting Ikusmen artifizialeko sistemek, eta orokorrean ikasketa automatikoko sistema guztiek, orokortze-ahalmen ona izatea ezinbestekoa da. Izan ere, askotan entrenamendu-multzoan emaitza onak lortzen dituzten sistemek eraginkortasun txikia erakusten dute entrenamendu-lagin hauetatik kanpo. Kasu batzuetan, neurona-sareek ebatzi behar duten zeregina burutzeko erabilgarriak diren ezaugarriak ikasi beharrean entrenamenduan zehar erabiltzen diren sarrera-irteera bikoteen laginak zuzenean memorizatzen dituzte, entrenamendu-zikloan galera-funtzioaren balioa txikituz baina entrenamendurako erabili ez diren test-

[4]Google Colab: https://colab.research.google.com

multzoko lagin berrietan emaitza txarrak lortuz. Fenomeno hau *overfitting* izenaz ezagutzen da [6], eta normalean sarea entrenatzeko eskuragai dagoen entrenamendu-multzoaren lagin-kopurua txikiegia delako edo diseinatu den arkitektura ebatzi nahi den zereginerako parametro gehiegi dituelako gertatzen da [24].

Hori dela eta, ikasketa sakoneko sistemak entrenatzeko datu-multzo egokiak erabiltzea itzelezko garrantzia du: neurona-sare batek emaitza hobeagoak lortzeko entrenamendu-zikloan laginen kopurua handitzea beti izango da onuragarria [6]. Lan honetako sareak bloke bakoitzean behin bakarrik entrenatu direnez, *overfitting* fenomenoa guztiz saihesten da: entrenamendu-zikloan sareek lagin bakoitza behin bakarrik ikusten dute.

4.4 Emaitzak



Irudia 4.4: *U-Net-Seg* sareak erabiliz objektuen eta estalduren segmentazio semantikoan lortutako irteeren adibideak.

U-Net-Seg sareei sarrera moduan ematen zaien kolorezko RGB irudia, neurona-sareek sarrerako iruditik abiatuz sortzen dituzten objektuen eta estalduren segmentazio-mapak eta benetako segmentazio-mapak 4.4 Irudian adierazten dira. 4.1 Atalean aipatu den bezala, objektuak eta estalkiak segmentatzeko bi neurona-sare desberdin entrenatu dira, bakoitza elementu bisual bakarrean espezializatu ahal izateko. Bi sare hauek arkitektura berdina (*U-Net-Seg*) eta parametroen kopuru bera dituzte, hau da, guztiz identikoak dira. Hala ere, nahiz eta sareen arkitektura berdinak izan, entrenamendu-zikloan helburu desberdinak izan dituzte: sare batek irudi batean objektuak detektatzeko eta ondoren hauek semantikoki segmentatzeko beharrezkoak diren pisuak lortu ditu eta besteak, berriz, prozesu bera estaldura dinamikoekin gauzatzeko behar diren parametroak eskuratu ditu. Hortaz, neurona-sareen arkitektura berdina hainbat zeregin desberdin burutzeko arazorik gabe berrerabili daiteke, entrenamendu-zikloan erabiltzen diren sarrera-irteera bikoteen laginak era egokian definitu ezkero.

4.4 Irudian adierazten diren hiru adibideetan *U-Net-Seg* sareen errendimendua orokorrean nahiko ona dela ikusi daiteke. Sareek kolore biziko objektuak (3. adibidea) erraztasun handiz segmentatzen dituzte gehienetan, baina beste kasu jakin batzuetan zailtasunak izan ahal dituzte: ingurunea islatzen duten metalezko gainazala (1. adibidea) duten objektuekin lan egiterako orduan akatsak agertu ahal dira, adibidez. Formari dagokionez, sareek objektu oso txikien definizio-maila mantentzeko arazoak izan ohi dituzte (1. eta 2. adibideak), baina orokorrean objektuen posizio zehatza eta gutxi gorabeherako forma arrakastaz detektatzeko gai dira. Hala eta guztiz ere, gizakiontzat zailak izan daitezkeen kasuetan ere (kolore iluneko objektuak, ondoko ingurunearen antzeko gainazalak dituzten objektuak, itzalekin nahasten diren objektuak, partzialki estalitako objektuak, objektu oso txikiak etab.) entrenatutako bi sareek errendimendu ona erakusten dute gehienetan (1. eta 2. adibideak), salbuespenak salbuespen. Gainera, neurona-sare bakoitzak helburu bezala duen elementua soilik zuzentasunez segmentatzeko gai da, hau da, sareek objektuak eta estaldura dinamikoak ez dituzte nahasten, nahiz eta batzuetan bi elementu bisual hauek oso antzekoak diren.

Neurona-sareak aurreko adibideen baliokideak diren irudi eta objektuekin entrenatu dira, baina 4.4 Irudian azaltzen diren sarrerako irudi zehatz hauek neurona-sareek lehenengo aldiz ikusi dituzte iragarpenak egiterako momentuan. Hau da, adibide bezala erakutsi diren kasuak test-multzo batetik atera dira, eta ez dira neurona-sareen entrenamendu-zikloan erabili. Horrela, irudien segmentazio semantikoa egiteko entrenatu diren bi *U-Net-Seg* neurona-sareen orokortze-ahalmena ona dela behatu daiteke, berriak diren laginetan ere errendimendu ona erakusten dute eta. Lehen aipatu den moduan, ikasketa sakonean oinarritutako ikusmen artifizialeko sistema batek orokortze-ahalmena garatzea ezinbestekoa da, hau datu berriekin lan egiterakoan emaitza onak lortzeko baldintza da eta.

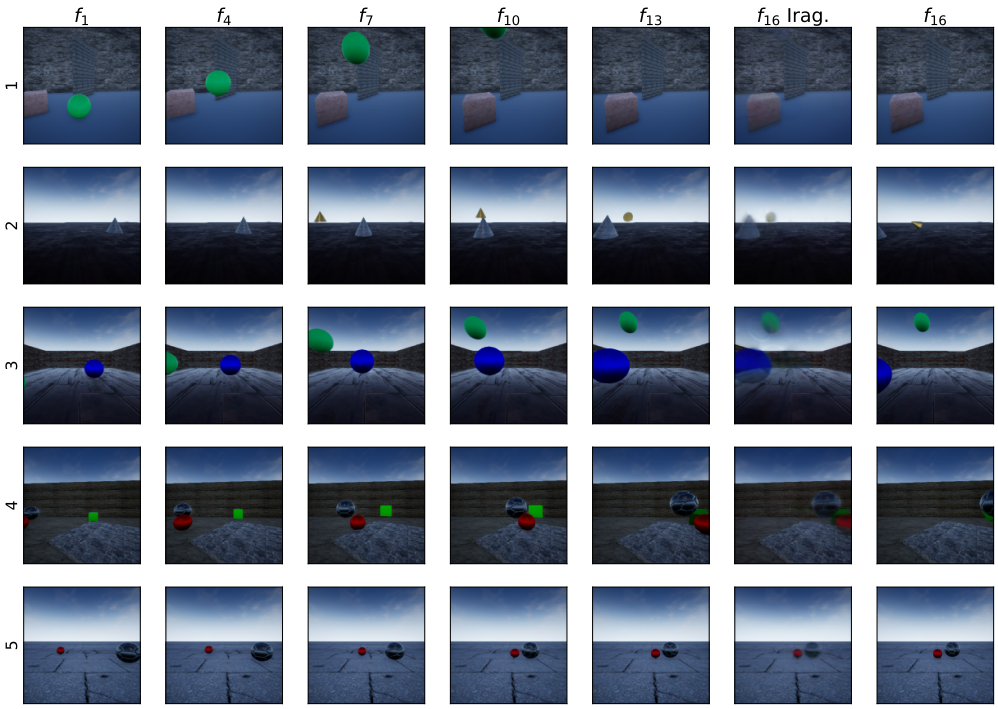
5 Etorkizuneko fotogramen iragarpena

Etorkizuneko fotogramen iragarpenaren zereginaren helburua bideo baten hainbat fotogramen (bideoa osatzen duten irudiak) sekuentzia sarrera moduan hartzen dituen ikusmen artifizialeko sistema batek sarrerako segida horren denborazko jarraipena izan daitekeen etorkizuneko fotograma sortzeko gaitasuna garatzea da. Beraz, sistema hauek bideotan agertzen den denborazko informazio bisuala erabiliz etorkizunean bideoan agertu ahal diren irudiak sortu behar dituzte [28]. Hori dela eta, etorkizuneko fotogramen iragarpenaren zeregina ikusmen artifizialeko sistemek etorkizuneko gertakizunei buruz arrazoitzeko duten gaitasun-maila neurtzeko tresna izan daiteke [17]. Etorkizuneko fotograma baten iragarpena ikasketa sakonean azpi-arlo berria da, neurona-sareak bideoak erabiliz entrenatuz ebatzi ahal diren problema desberdinak aztertzen dituena, eta azken urteetan aurrerapen ugari lortu dira.

Etorkizuneko fotograma bat aurresateko gai den sistema batek bideoak osatzen dituzten irudien eduki kontzeptuala eta eduki hauek duten denborazko eboluzioa modelizatzeko ahalmena izan behar du, hau da, bideoan agertzen diren elementu bisualek erakusten duten zinematika eta dinamika modu intuitibo batean ulertzeko gai izan behar dira [16]. 3D-ko espazio bat adierazten duten kolorezko RGB irudietatik abiatuz (*IntPhys*) erlazio konplexu hauek

matematikoki eskuz definitzea oso konplexua izango litzateke eta, beraz, kasu honetan ikasketa automatikoa erabiltzea ezinbestekoa da.

Zeregin honetan lan egiten hasterakoan, burura datorren lehenengo ideia zuzenean prozesatu gabeko kolorezko RGB irudiak entrenamendu-zikloan sarrera-irteera lagin bezala erabiltzea izan ohi da. Izan ere, neurona-sareak entrenatzeko mota honetako datu pilo bat oso erraz eskuragarri daude: YouTube-en igota dauden bideo gehienetan hainbat motako mugimenduak eta objektu desberdinen dinamika eta zinematika adierazita agertzen dira. Hala eta guztiz ere, nahiz eta entrenamendu-zikloan erabili ahal diren datuen kantitatea itzela izan, neurona-sareek zeregin zehatz hau burutzeko arazo handiak izan ohi dituzte. Beraz, ikasketa sakonaren arloan bideokin lan egitean etorkizuneko fotogramen iragarpenak eraginkortasunez egitea gaur egun ere guztiz ebatzi gabe dagoen problema da [28].



Irudia 5.1: Sarrera moduan zuzenean kolorezko RGB irudiak erabiliz etorkizuneko fotograma baten iragarpena burutzerakoan *U-Net-Ira* sareak lortutako irteeren adibideak.

Gogoratu beharra dago neurona-sareek irudiak tentsore moduan antzematen dituztela: galera-funtzioaren balioa minimizatzerako orduan sareentzat pixel guztiak garrantzi berdina dute. Neurona-sareek ez dituzte objektuak, koloreak, testurak eta antzeko kontzeptuak berez antzematen; tentsoreen elementuak diren zenbakiak ikusten dituzte bakarrik. Hori dela eta, mugimenduan dagoen objektuak fotograma batzuen azalera osoaren %5a soilik betetzen badu, sareak denborazko iragarpena egiterako orduan sarrerako irudietan geldi dagoen atzealdea

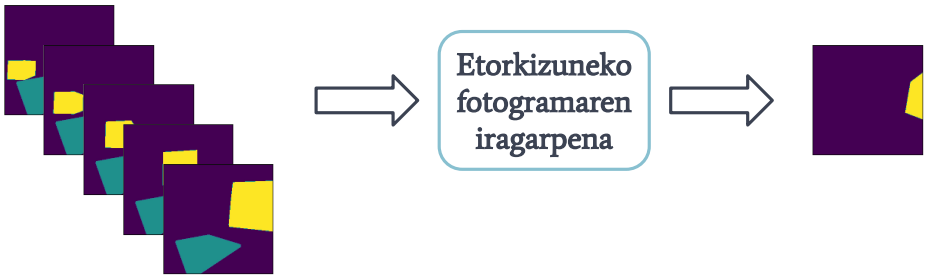
zuzenean kopiatuz $\sim$ %95ko zuzentasuna lortzen duela oso azkar ikasiko du, galera-funtzioa era honetan azkar txikitzen da eta. Hortaz, neurona-sareak sarrera bezala hartzen duen fotogramen sekuentzian agertzen den atzealde geldikorra ondo birsortuz emaitza onak lortzen dituela barneratuko du, eta mugimenduan dauden elementu dinamikoen ibilbidea eta etorkizuneko forma aurresaten ikastea galera-funtzioaren testuinguruan etekin txikiko eta zailtasun handiko erronka denez, guztiz baztertuko du zeregin hau; nahiz eta hori objektu makroskopikoen fisika ikasteko denborazko iragarpenaren zatirik interesgarria izan. Hori dela eta, ikasketa-prozesua gertatzen denean sarearen arreta bereziki mugimenduan dauden objektuetan jartzeko modua aurkitzea beharrezkoa da.

Oztopo hau gainditzeko, neurona-sareak arrakasta neurtzeko duen modua aldatu behar da. Kolorezko RGB irudietan gordeta dagoen informazio asko (argitasun-maila, objektuen koloreak, gainazalen testurak, hodeien itxura etab.) objektu makroskopikoen zinematika eta dinamika ikasteko zereginean arazo handirik gabe arbuiatu daiteke. Beraz, aurreko atalean landu diren segmentazio-mapak erabiliz bideoetako eszena bisualen goi-mailako informazio kontzeptuala kodifikatu daiteke, horrela sarearen arreta osoa objektuen posizioaren denborazko eboluzioan eta objektuen formen aldaketan jarriz. Bestalde, kolorezko RGB irudietatik abiatuz beharrezkoak diren elementu dinamikoen segmentazio-mapa hauek lortzeko, lehen entrenatu diren bi *U-Net-Seg* sareen irteerak konbinatuko dira.

5.1 Helburua

5.1 Irudiaren bidez frogatu den bezala, zuzenean kolorezko RGB irudiekin lan egiteak etorkizuneko fotogramaren iragarpena gehiegi zailtzen du. Aplikazio askotan, berriz, abstrakzio-maila handiagoko informazioa adierazten duten segmentazio-mapak erabiltzea nahikoa izan daiteke eszena bisual batean agertzen diren elementu dinamikoen denborazko iragarpenak egiterako orduan [17]. Hori dela eta, lan honetan objektuen eta estalduren segmentazio-mapak konbinatzen dituzten fotogramen segidetan agertzen diren objektu makroskopikoen dinamikak eta antzeko fenomeno fisikoak modelizatzen dituen ikusmen artifizialeko sistema lortu nahi da.

Etorkizuneko fotograma iragartzeko entrenatuko den neurona-sareak sarrera moduan 5 fotogramaz osatutako sekuentzia hartzen du (objektuen eta estalduren segmentazio-mapen konbinazioak), haien artean 3 fotogramako tartea utziz, eta irteeran segida honen azkenengo fotogramatik abiatuz 3 fotograma aurrerago agertuko den objektuen segmentazio-mapa bueltatzea du helburu bezala (5.2 Irudia). Beraz, beste modu batean esanda, sarea entrenatzeko fotogramen segidek eta etorkizuneko fotogramek osatzen dituzten $(f_{t-12}, f_{t-9}, f_{t-6}, f_{t-3}, f_t) \rightarrow f_{t+3}$ moduko bikoteak entrenamendu-lagin gisa erabiliko dira.



Irudia 5.2: *U-Net-Ira* sarearen helburua.

5.2 Sarearen arkitektura

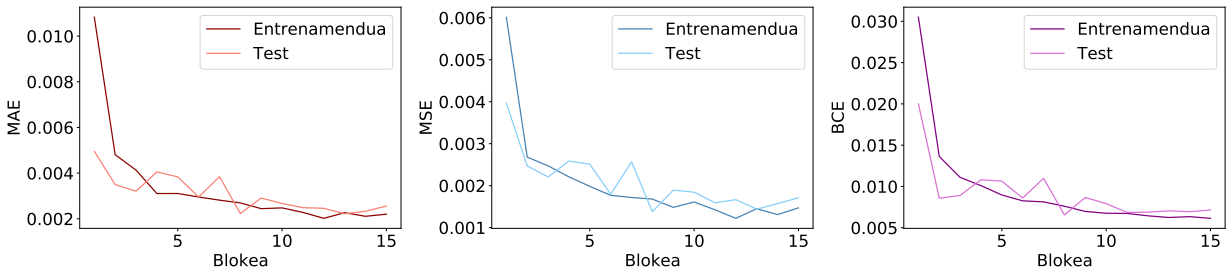
Segmentazio semantikorako arrakastaz erabili den *U-Net-Seg* sareari aldaketa batzuk eginez etorkizuneko fotogramaren iragarpenaren zereginean antzeko arkitektura berrerabiltzea posiblea da. Izan ere, *U-Net* arkitekturan oinarritutako neurona-sareek beste sare-diseinu mota batzuk baino emaitza hobeagoak lortu dituzte etorkizuneko fotogramen iragarpenean, zehazki segurtasun-kameren bideoetan anomaliak detektatzeko balio duten ikusmen artifizialeko sistemen diseinuan [14]. Aurrekoa kontuan hartuz, mota desberdineko zereginak burutzerako orduan ikasketa sakoneko neurona-sareek erakusten duten moldagarritasun-ahalmen handia nabaria da.

Etorkizuneko fotogramaren iragarpena arrakastaz burutu ahal izateko, segmentazio semantikorako erabili den neurona-sarearen modelizazio-ahalmena handitzea beharrezkoa da. Beraz, *U-Net* arkitekturaren jatorrizko ideia (4.2 Irudia) mantenduz kodetzailearen eta deskodetzailearen bloke bakoitzean beste 3×3-ko konboluzio-geruza bat gehitu da eta hasierako blokea 16 iragazkiekin hasi beharrean 64 iragazkiekin hasten da (sarearen maila bakoitzean iragazkien kopurua bikoizten dela gogoratu behar da). Bi aldaketa hauen bidez neurona-sareak doitu ahal dituen parametroen kopurua handitzen da eta, ondorioz, sareak sarrerako datuen abstrakzio handiagoko barne-errepresentazio erabilgarriak lortzeko duen gaitasuna ere. Etorkizuneko fotogramaren iragarpenerako erabiliko den neurona-sarearen arkitektura berria *U-Net-Ira* izena jaso du.

5.3 Entrenamendua

Berriro ere, *IntPhys* datu-multzo osoa 15 bloke desberdinetan banandu da. Bestalde, bloke bakoitzean dauden bideoak beste bi multzotan banantzen dira ere: 900 bideo entrenamendu-zikloan erabiliko den entrenamendu-multzoa osatzen dute eta geratzen diren beste 100 bideoak test-multzorako gordetzen dira. Segmentazio semantikoa burutzeko erabili diren neurona-sareekin egin den bezala, entrenamendu-zikloaren luzera murrizteko sarrerako fotogramen bereizmena erdira jaisten da, (288×288)-tik (144×144)-ra pasatuz.

Aurreko kasuaren moduan, neurona-sarea entrenatzeko *Adam* optimizatzailea [12] erabili da, $\epsilon = 0.001$-ko ikasketa-erritmoarekin. *U-Net-Ira* bloke bakoitzean behin bakarrik entrenatzen da ere, *overfitting* fenomenoa ekiditeko, eta sareak blokeen entrenamendu-zikloaren amaieran lortzen dituen pisuen balioak gorde eta hurrengo blokean entrenamendua hasi baino lehen berriro ezartzen dira. Neurona-sareak ~ 30 minutu behar ditu bloke bakoitzean entrenamendu-zikloa burutzeko (sarea entrenatzeko guztira 7 ordu eta 30 minutu behar dira, beraz) eta guztira 10.849.793 parametro desberdin ditu. Entrenamendu-zikloan 2.3.3 Atalean definitu diren 3 galera-funtzioak erabili dira: $\mathcal{L}_{BCE}$, $\mathcal{L}_{MSE}$ eta $\mathcal{L}_{MAE}$. Horrela, etorkizuneko fotogramaren iragarpenaren zeregina burutzeko entrenatu den *U-Net-Ira* sarearen irteeran galera-funtzioaren aukerak duen eragina behatu daiteke. 5.3 Irudian galera-funtzio desberdinen balioa entrenamenduan zehar eta bloke bakoitzaren test-multzoan adierazten da.

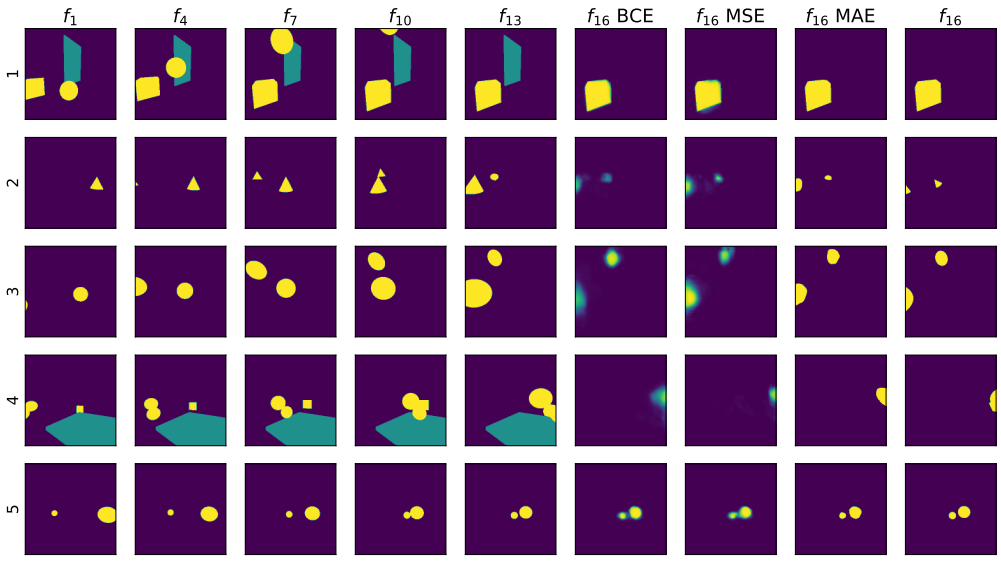


Irudia 5.3: $\mathcal{L}_{BCE}$, $\mathcal{L}_{MSE}$ eta $\mathcal{L}_{MAE}$ galera-funtzioekin entrenatutako *U-Net-Ira* neurona-sareek entrenamendu-zikloan zehar eta bloke bakoitzarekin lotutako test-multzoan lortzen dituzten galera-balioak.

5.4 Emaitzak

U-Net-Ira neurona-sarean oinarritutako ikusmen artifizialeko sistemak objektuen ibilbideak eta zinematika sinplea ulertzeko gaitasuna duela 5.4 Irudian adierazten diren adibideetan behatu daiteke. Zuzenean kolorezko RGB irudiekin lan egitean lortzen diren iragarpenekin alderatuz (5.1 Irudia), segmentazio-mapak erabiltzean neurona-sareak bueltatzen dituen irteeretan *U-Net-Ira* sareak objektu makroskopikoen zinematika eta dinamika modelizatzeko duen gaitasuna asko handitu dela argi dago, batez ere objektuen etorkizuneko posizioaren zehaztasunari dagokionez (alderaketa errazteko, 5.4 Irudian eta 5.1 Irudian agertzen diren adibideak berdinak dira). Izan ere, lehen aipatu den bezala, neurona-sare batek bideo baten fotogramen denborazko eboluzioa zehaztasunez iragartzeko bideoan agertzen diren elementu bisualek (kasu honetan objektuen eta estalduren segmentazioak) jarraitzen dituzten lege fisikoak modelizatu behar ditu [16].

5.4 Irudian *U-Net-Ira* sareak objektu geldikor baten denborazko eboluzioa aurresatearen problema tribialean arazorik ez duela ikusten da (1. adibidea). Hala ere, segmentazio semantikoaren kasuan gertatzen zen bezala, neurona-sareak mugimenduan dauden objektu txikien forma zehatzak denboran zehar mantentzeko arazoak dituela argi geratzen da ere (2. adibidea). Bestalde, *U-Net-Ira* sareak objektu dinamikoak irudien ertzetatik desagertu ahal



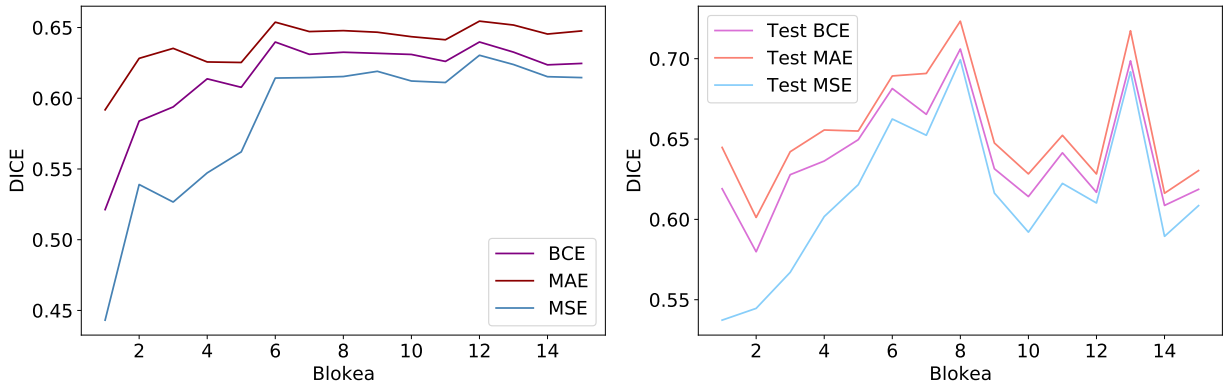
Irudia 5.4: Etorkizuneko fotogramaren iragarpena burutzen duen neurona-sarea galera-funtzio desberdinekin lortzen diren irteeren adibideak.

direla ulertzen du eta hainbat objektu aldi berean jarraitzeko ahalmena du (2-4 adibideak). Azkenik, objektuen arteko talketan gertatzen diren norabide aldaketak iragartzeko gaitasuna ere duela behatu daiteke (5. adibidea).

Segmentazio-mapez osatutako $(f_{t-12}, f_{t-9}, f_{t-6}, f_{t-3}, f_t) \rightarrow f_{t+3}$ motako sarrera-irteera bikoteak entrenamendu-lagin bezala erabiliz, *U-Net-Ira* zehaztasun gehiagorekin objektuen ibilbideak aurresaten ditu eta, ondorioz, ibilbide hauek gobernatzen dituzten zinematika eta dinamika intuitiboki hobeto ikasten ditu ere. Beraz, entrenamendu-zikloan etorkizuneko fotogramen objektuen segmentazio-mapak helburu bezala erabiliz, neurona-sarearen arreta guztia objektu dinamikoen denborazko eboluzioaren iragarpenean jartzen da, hau baita kasu honetan galera-funtzioa txikitzeko bide bakarra.

Hala ere, 5.1 Irudian neurona-sareak objektuen formak eta tamaina zehatzak denboran zehar mantentzeko zailtasunak dituela behatu daiteke, bereziki $\mathcal{L}_{BCE}$ eta $\mathcal{L}_{MSE}$ galera-funtzioekin entrenatu diren sareen kasuan. Argi ikusi daiteke $\mathcal{L}_{MAE}$ galera-funtzioaren bidez entrenatu den neurona-sareak irteera kualitatiboki hobeagoak lortzen dituela: definizio-maila handiagoko iragarpenak sortzen ditu. 5.5 Irudian entrenamendu-zikloan zehar eta test-multzoetan galera-funtzio zehatzekin entrenatu diren sareek lortu duten *Dice* koefizientearen balioa adierazten da. 5.5 Irudia kontuan hartuz, etorkizuneko fotogramaren iragarpena segmentazio-mapak erabiliz burutzen duen neurona-sare bat entrenatzeko, lan honetan proposatu diren hiru galera-funtzioen artean kuantitatiboki emaitza hoberenak $\mathcal{L}_{MAE}$ galera-funtzioak lortzen ditu. Horrela, problema zehatz bat ebazteko erabiliko den neurona-sare bat entrenatzerako orduan galera-funtzioaren aukera garrantzi handiko diseinu-erabakia dela frogatzen da.

Berriro ere, 5.1 Irudian agertzen diren adibideak test-multzo batetik atera diren fotogramen segiden bidez sortu direla azpimarratu nahi da. Hori dela eta, *U-Net-Ira* orokortze-ahalmen nahiko ona duela ikusi daiteke eta adibide berri hauetan entrenamendu-zikloan zehar ikasitako modelizazio fisikoaren bidez fisikoki onargarriak diren iragarpen berriak egiteko gaitasuna duela behatzen da.



Irudia 5.5: $\mathcal{L}_{BCE}$, $\mathcal{L}_{MSE}$ eta $\mathcal{L}_{MAE}$ galera-funtzioekin entrenatutako *U-Net-Ira* neurona-sareek entrenamendu-zikloan zehar eta bloke bakoitzarekin lotutako test-multzoan lortzen dituzten *Dice* koefizientearen balioak.

6 Tresnak

Lan honetan aztertu diren neurona-sare guztiak *TensorFlow*[5] eta *Keras*[6] programa-liburutegiak erabiliz diseinatu, programatu eta entrenatu dira. *Keras* bereziki lagungarria izan da, ikasketa sakonaren esparruan ohikoak diren geruza asko jada definituta eta erabiltzeko prest ditu eta. Gainera, neurona-sareak entrenatzerako orduan erabilgarriak diren funtzio eta metodo ugari eskaintzen ditu, eta sareen entrenamendu-zikloaren oinarria den *Stochastic Gradient Descent* algoritmoa inplementatuta dago. Bestalde, gehien erabiltzen diren galera-funtzioak definituta daude ere. Hau guztia dela eta, *Keras* programa-liburutegiak ikasketa sakoneko neurona-sareekin lan egitearen zailtasun ugari errazten ditu eta askotan nahiko konplexua izan daitekeen adimen artifizialaren arloa demokratizatzeko balio duen oso tresna garrantzitsua da.

Sareak entrenatzeko *Google Colab*[7] ingurunea erabili da proiektu honetan, neurona-sareen entrenamenduetan GPU unitate bat erabiltzeko aukera ematen duelako. Irudiekin lan egiten duten ikasketa sakoneko sistemak entrenatzeko GPU bat erabiltzea ezinbestekoa dela argi geratu zen hasieratik: CPU on batean 16 ordu behar zituen entrenamendu-zikloa *Google Colab*-eko GPUa erabiliz $\sim$ 20 minututan burutzen da.

[5]TensorFlow: https://www.tensorflow.org
[6]Keras: https://keras.io
[7]Google Colab: https://colab.research.google.com

Kodea Proiektu osoaren inplementazioa ondorengo biltegi digitalean eskuragarri dago: https://github.com/jperezvisaires/tfg-intphys

7 Ondorioak

Bideo baten hainbat fotogramez osatutako sekuentzia batetik abiatuz etorkizuneko fotograma bat iragartzeko gaitasuna duen ikusmen artifizialeko sistema bat eraikitzea ez da zeregina erraza izan. Hasiera batean, zuzenean prozesatu gabeko kolorezko RGB irudiekin lan egiteko ahalegin asko egin ziren, baina arrakastarik gabeko saiakera anitz egin ondoren kasu zehatz honetan etorkizuneko fotograma aurresateko era ezegokia zela argi geratu zen. Izan ere, neurona-sareak atzealde geldikorra ondo berregiten zuen arren, objektu mugikorren ibilbideak eta haien etorkizuneko posizioak iragartzeko gaitasuna ez zuen batere garatzen (5.1 Irudia): zailtasun-maila handiagoko zeregin hau ikasteak sarearen galera-funtzioaren minimizazioan eragin arbuiagarriegia zuen.

Hori dela eta, neurona-sareak objektuen zinematika eta dinamika ikasi ahal izateko, irudietan kodifikatuta dagoen informazioa era adierazgarriago batean jaso behar du. Sistemak burutu behar dituen iragarpenen zailtasun-maila errazteko helburuarekin, sareak jasotzen dituen datuetan abstrakzio-maila handiago bat erabiltzea beharrezkoa da: objektuen eta estalduren forma eta posizioa soilik adierazten dituzten segmentazio-mapak baliagarriak izan daitezke hau lortzeko [17]. Segmentazio-mapak erabiltzean, neurona-sareak objektuen zinematika eta dinamika ikasteko erabilgarria ez den informazio asko baztertu dezake eta ikasketa-prozesuan zehar bere arreta osoa benetan ebatzi nahi dugun problemarako garrantzitsua den objektuen denborazko eboluzioan aplikatu dezake. Horrela, sareak lege fisikoak ulertzeko baliagarriak ez diren koloreak, gainazal mota desberdinak, argitasun-maila eta antzekoak diren propietateak alde batera utzi ahal ditu, zinematika eta dinamika ikasteko erabilgarriak diren ezaugarriak soilik mantenduz.

Emaitzetan ikusi daitekeen bezala, ikusmen artifizialeko sistemak objektuen etorkizuneko posizioak aurresateko nolabaiteko gaitasuna garatzea lortu du (5.4 Irudia), baina denboran zehar objektuek dituzten forma zehatzak zehaztasun handiz mantentzeko zailtasunak ditu. Normala den bezala, forma esferikoak dituzten objektuekin arazo gutxiago ditu, baina kuboak edo konoak diren objektuen definizio-maila kasu gehienetan iragarpenetan jaisten da, objektu mota horiek dituzten ertz zorrotzak direla eta. Hala eta guztiz ere, objektu mugikorren ibilbideak iragartzeko lortutako ahalmen hau neurona-sareak zinematikari eta dinamikari buruzko nolabaiteko ulermen fisiko intuitiboa garatu duela adierazten du, ulermen hori zeregin hau modu egokian ebazteko baldintza beharrezkoa baita.

Bestalde, gogoan izan behar da lan honetan landu den sistemak ikasketa-prozesua burutzerako orduan izan duen informazio-kopurua mugatua izan dela, adin txikiko ume batekin konparatu ezkero. Izan ere, umeek haien ingurunearekin interakzionatzeko aukera dute eta zinematikaren eta dinamikaren arloetan agertzen den kausalitatea hobeto ulertzeko oso tresna baliagarria da hau. Proiektu honetan aztertu den neurona-sareak, berriz, fisikoki posibleak diren bideo

sintetikoen kopuru handia behatuz ikasi du bakarrik, bere ingurunean inolako eragin fisikorik izateko gaitasunik gabe. Sistema adimentsuek etorkizuneko gertakizunei buruz egiten dituzten iragarpenak hobetzeko haien ingurunearekin interakzionatzeko ahalmena izatea garrantzitsua da, eta norabide honetan ikerkuntza-ahalegin handiak egin dira azken urteetan [5].

Nahiz eta *U-Net-Ira* neurona-sareak lortu dituen emaitzak nahiko onak izan diren, lan honetan diseinatu den ikusmen artifizialeko sistemak muga nabariak ditu. Etorkizun hurbileko gertakizunak aurresan dezake bakarrik eta memoria txikia du ($(f_{t-12}, f_{t-9}, f_{t-6}, f_{t-3}, f_t) \rightarrow f_{t+3}$); epe luzera iragarpenak egiteko gaitasuna garatzeko behar den barne-errepresentazio fisikoa eskuratzea problema askoz konplexuagoa da. Neurona-sareen denborazko iragarpen-ahalmenak hobetzeko helburuarekin, lan batzuk denbora-segidekin lan egiteko sortu ziren *Long Short-Term Memory* (LSTM) geruzak ohiko konboluzio-geruzekin konbinatzen dituzte [5][28].

Azkenik, lan honetan etorkizuneko fotograma baten iragarpena egiteko $\mathcal{L}_{MAE}$ galera-funtzioa erabiltzean kalitate hobeko irteerak lortzen direla behatu da (5.5 Irudia). Hala ere, iragarpenetan agertzen diren formen definizio-maila are gehiago hobetzeko, galera-funtzio bezala *Generative Adversarial Network* (GAN) [7] bat erabili da beste lan batzuetan [16][14]. Sare-diseinu mota hauen atzean dagoen oinarrizko ideia hurrengoa da: iragarpenak egiten dituen sarearen irteeren benetakotasuna neurtzeko helburuarekin entrenatzen den beste neurona-sare bat galera-funtzio bezala erabiltzea, hau da, galera-funtzioa ere ikasketa-prozesuan zehar hobetzea.

Erreferentziak

[1] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 2013.

[2] J. Brownlee. How to choose loss functions when training deep learning neural networks, 2019. URL: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/.

[3] F. Chollet. *Deep learning with Python*. Manning Publications Co, 2018.

[4] J. Dzieza. How hard will the robots make us work?, 2020. URL: https://www.theverge.com/2020/2/27/21155254/automation-robots-unemployment-jobs-vs-human-google-amazon.

[5] C. Finn, I. J. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. *CoRR*, 2016.

[6] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press. URL: http://www.deeplearningbook.org.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. *ArXiv*, 2014.

[8] U. Goswami, editor. *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell, 2010.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, 2017.

[10] J. Jordan. Convolutional neural networks, 2017. URL: https://www.jeremyjordan.me/convolutional-neural-networks/.

[11] J. Jordan. An overview of semantic image segmentation, 2018. URL: https://www.jeremyjordan.me/semantic-segmentation.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, 2015.

[13] L. Nieradzik. Loss functions for segmentation, 2018. URL: https://lars76.github.io/2018/09/27/loss-functions-for-segmentation.html.

[14] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection - A new baseline. *CoRR*, 2017.

[15] D. Marr. *Vision : a computational investigation into the human representation and processing of visual information*. MIT Press, 2010.

[16] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, 2016.

[17] N. Neverova, P. Luc, C. Couprie, J. J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. *CoRR*, 2017.

[18] S. Narayanan. An update about face recognition on Facebook, 2019. URL: https://about.fb.com/news/2019/09/update-face-recognition/.

[19] J. Piaget. *The psychology of the child*. Basic Books, 1969.

[20] D. L. Poole and A. K. Mackworth. *Artificial intelligence : foundations of computational agents*. Cambridge University Press, 2010.

[21] D. L. Poole, A. K. Mackworth, and R. Goebel. *Computational intelligence : a logical approach*. Oxford University Press, 1998.

[22] R. Riochet, M. Ynocente Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *CoRR*, 2018.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015.

[24] A. Rosebrock, editor. *Deep Learning for Computer Vision with Python*. PyImageSearch, 2017.

[25] C. E. Shannon. XXII. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1950.

[26] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, 2016.

[27] R. Szeliski. *Computer vision : algorithms and applications*. Springer, 2011.

[28] Y. Zhou, H. Dong, and A. El Saddik. Deep learning in next-frame prediction: A benchmark review. *IEEE Access*, 2020.