Modelarea distorsiunilor de imagine induse de refracție în camerele din spatele obiectelor transparente

Rezumatul tezei de doctorat

Szabolcs Pável

Facultatea de Matematică și Informatică

Universitatea Babeş-Bolyai, Cluj-Napoca

Conducător științific

Prof. Dr. Horia F. Pop

2024

Abstract

Camerele video sunt utilizate frecvent în sistemele avansate de asistență a conducătorului auto. Cea mai frecventă poziție de montare a camerelor frontale este lângă oglinda retrovizoare. Deoarece lumina este refractată la suprafața parbrizului, aceasta acționează ca un element optic și provoacă distorsiuni complexe ale imaginilor. Aceasta este o problemă pentru algoritmii de viziune computerizată, deoarece aceștia presupun un model precis al camerei.

După prezentarea conceptelor fundamentale ale modelelor de camere și ale metodelor de învățare profundă, sunt prezentate două soluții pentru problema distorsiunilor cauzate de un obiect transparent în calea optică.

În primul rând, este propus un model în care urmărim în mod explicit calea razelor de lumină. Atât componentele globale, cât și cele locale ale distorsiunilor sunt modelate cu ajutorul funcțiilor de bază radiale, iar pentru a găsi parametrii optimi ai modelului se utilizează un algoritm de calibrare bazat pe ținte de calibrare cu tablă de șah. Metoda este testată pe imagini reale capturate de o cameră plasată în spatele unui obiect de sticlă, precum și pe imagini distorsionate generate sintetic.

În al doilea rând, este prezentată o abordare bazată pe învățarea profundă, în care o rețea neuronală convoluțională este utilizată pentru a estima direct distorsiunile pe baza unei singure imagini. Similar cu prima abordare, sunt generate seturi de date sintetice și reale la scară largă pentru a antrena modelele. Rețeaua este antrenată utilizând funcții de eroare bazate pe reconstrucția imaginii, iar segmentarea semantică și fluxul optic sunt incluse ca sarcini auxiliare pentru îmbunătățirea rezultatelor.

Cuprins

Capitolul 1

Introducere

Sistemele avansate de asistență a conducătorului auto au fost adoptate pe scară largă în automobilele moderne. Aceste sisteme oferă funcții de siguranță, cum ar fi frânarea automată de urgență, avertizare la părăsirea benzii de circulație și asistență la menținerea benzii de circulație, recunoașterea semnelor de circulație, asistență inteligentă la viteză și multe altele. Aceste sisteme se bazează pe mai mulți senzori pentru a construi o reprezentare exactă a mediului din jurul mașinii. Aceste reprezentări trebuie să fie precise și robuste, deoarece aceste sisteme funcționează în medii de siguranță critice. Printre senzorii frecvent utilizați se numără camerele foto, radarul, ultrasunetele și, în unele cazuri, lidarul. Dintre acești senzori, camerele sunt cele mai versatile. Acestea furnizează informații de înaltă rezoluție la un cost redus, permițând extragerea unei semantici bogate despre mediul înconjurător. Întrucât infrastructura rutieră umană este construită în cea mai mare parte pe baza vederii, pentru anumite sarcini, cum ar fi interpretarea marcajelor rutiere și a semnelor de circulație, camerele sunt singura soluție.

Atunci când o scenă este capturată cu ajutorul unei camere, se pierd informații despre structura lumii. Pentru a putea raționa cu privire la poziția obiectelor din jurul mașinii, aceste informații trebuie să fie recuperate. Acest lucru se poate realiza utilizând sisteme cu mai multe camere (stereo), în care, prin observarea aceluiași obiect din mai multe unghiuri, poziția acestuia poate fi recuperată utilizând triangulația. Atunci când se utilizează sisteme monoculare cu o singură cameră, se pot folosi în schimb imagini capturate la momente diferite și se poate reconstrui lumea folosind abordări de tip Structure-from-Motion. În ultimul timp, au fost propuse și tehnici de învățare profundă, care pot prezice informații privind adâncimea pe baza unei singure imagini. Indiferent de sistemul de camere și de tehnica utilizată, toate aceste

metode necesită o mapare precisă de la punctele 3D din lume la pixelii 2D din imagine - aceste mapări se numesc *modele de cameră*.

Modelul ideal al camerei depinde de proprietățile camerei în sine. Pentru camerele cu câmp vizual mic sunt suficiente modele simple bazate pe proiecția în perspectivă. Camerele cu unghi larg, cum ar fi camerele fisheye, deviază semnificativ de la modelul perspectivei, imaginile sunt *distorsionate* geometric, ceea ce trebuie luat în considerare. Aceste modele caracterizează proprietățile camerelor la nivel *global* - doar un număr mic de parametri sunt utilizați pentru a descrie maparea pe întregul câmp vizual. Modelele globale pot eșua, de exemplu, atunci când camerele sunt plasate în spatele unor obiecte transparente, care refractă razele de lumină incidente. Aceste obiecte pot introduce un comportament neliniar în funcția de mapare, care trebuie modelat în consecință - folosind modele *locale*.

Parametrii modelelor de camere sunt estimați în timpul *calibrării camerelor*. Metodele de calibrare pot fi împărțite în două categorii principale. Atunci când o cameră este instalată pentru prima dată, are loc o *calibrare inițială (offline)*. Acest proces se realizează folosind medii controlate, în prezența unor ținte de calibrare specifice (de exemplu, tablă de șah) și a unor dispozitive de măsurare. Metodele de calibrare inițială pot estima parametrii cu o precizie ridicată, dar necesită mult timp și sunt costisitoare. În timp ce camera este utilizată, aceasta se poate decalibra: din cauza efectelor mediului, cum ar fi schimbările mari de temperatură și stresul mecanic (de exemplu, vibrațiile), proprietățile lentilelor se modifică față de valorile inițiale. Pentru a corecta aceste probleme, poate fi aplicată o *autocalibrare online*. Metodele de autocalibrare ajustează camera în timp ce aceasta funcționează, fără a fi nevoie de ținte de calibrare specifice în mediu.

1.1 Obiective

Motivația acestei teze provine de la *camerele inteligente* utilizate pentru sistemele avansate de asistență a conducătorului auto. Camerele inteligente pentru automobile sunt dispozitive compacte, care includ atât sistemul optic, cât și unitățile de procesare (utilizând System-on-Chip) pentru implementarea funcțiilor de siguranță. Aceste dispozitive sunt montate în zona de deasupra oglinzii retrovizoare a mașinii, în spatele parbrizului din față al mașinii. Acest lucru vine cu provocări unice: parbrizul refractiv aflat în imediata apropiere a camerei *distorsionează* puternic imaginile. Aceste distorsiuni sunt foarte neliniare, iar modelele standard de camere

nu le pot descrie cu precizie. Lucrarea prezentată în această teză se concentrează pe modelarea acestor distorsiuni: este propusă atât o metodă de calibrare inițială bazată pe modelarea explicită a refracției luminii la nivelul parbrizului, cât și o metodă de autocalibrare bazată pe învățarea profundă.

Primul obiectiv al acestei teze este de a propune un algoritm de calibrare inițială care să îndeplinească următoarele cerințe:

- Modelul trebuie să fie capabil să descrie atât componentele globale, cât și cele locale ale distorsiunilor neliniare ale parbrizului. Acest lucru permite luarea în considerare atât a formei globale a parbrizului, precum și neregularitățile locale ale suprafeței.
- Modelul trebuie să se bazeze pe fizică: razele de lumină care trec prin mediul transparent trebuie să fie trasate când sunt refractate pe suprafețe. Această abordare oferă o modalitate ușoară de a încorpora cunoștințele prealabile privind proprietățile generale ale parbrizului și, de asemenea, oferă posibilitatea de a analiza mai detaliat distorsiunile.
- Progresele recente ale framework-urilor de diferențiere automată trebuie valorificate, deoarece oferă posibilitatea de a construi și optimiza modele complexe. Acest lucru oferă o mare flexibilitate în selectarea modelelor pe care dorim să le utilizăm.

În plus față de metoda de calibrare inițială, al doilea obiectiv al acestei teze este de a propune o metodă de autocalibrare, cu cerințele:

- Metoda de autocalibrare trebuie să se bazeze pe tehnici de învățare profundă. Modelele de învățare profundă pot fi ușor integrate în sistemele de camere auto, deoarece acestea oferă acceleratoare specifice pentru sarcinile de lucru ale rețelelor neuronale.
- Similar metodei de calibrare inițială, rețeaua neuronală trebuie să utilizeze un model de distorsiune care să poată reprezenta atât deformările globale, cât și cele locale ale imaginii.
- Pentru formarea rețelelor neuronale, trebuie construit un set de date bazat pe măsurători reale ale distorsiunilor parbrizului. Setul de date ar trebui să conțină imagini sintetice și reale, iar transferabilitatea între datele simulate și cele reale ar trebui să fie analizată. Acest lucru este necesar, deoarece seturile de date la scară largă pentru calibrarea camerelor nu sunt disponibile publicului.

- Funcțiile de eroare bazate pe reconstrucția imaginilor sunt preferate celor care se bazează pe date de distorsiune reale. Aceste funcții de eroare deschid posibilitatea de a extinde metoda pentru a utiliza seturi de date fără etichete reale.

- Sarcinile auxiliare ar trebui integrate în arhitectura rețelei și ar trebui analizate efectele de performanță ale acestei abordări multitask. Principalele sarcini auxiliare candidate sunt segmentarea semantică și fluxul optic, deoarece aceste predicții sunt deja disponibile în majoritatea rețelelor neuronale din domeniul auto.

1.2 Contribuții

Contribuțiile acestei teze sunt următoarele:

- Sunt propuse noi soluții pentru cazurile de utilizare în care camera este montată în spatele unui obiect transparent, care introduce distorsiuni. Contribuția are impact asupra sistemelor avansate de asistență a șoferului, în care camera este montată deasupra oglinzii retrovizoare a mașinii. Sunt propuse atât o calibrare inițială (cap. 4, cap. 5), cât și o metodă de autocalibrare (cap. 6).

- Se propune o metodă de calibrare inițială pentru modelarea suprafeței neuniforme a unui obiect transparent (cap. 4).) Se presupune că forma globală a obiectului este cunoscută, ceea ce este adesea cazul în industria auto, în timp ce componentele locale ale distorsiunii sunt descrise utilizând un model parametric. Deoarece forma globală este cunoscută, alegem abordarea de a modela în mod explicit refracția luminii, spre deosebire de abordarea mai explorată a modelelor generalizate de cameră.

- Pentru a testa metoda propusă, un set de date cu imagini distorsionate este înregistrat utilizând o cameră Raspberry Pi montată în spatele unui obiect din sticlă curbată. În plus, se utilizează un set de date sintetice pentru a furniza distorsiunile reale disponibile.

- Teza oferă o analiză a distorsiunilor parbrizului bazată pe modelul generativ propus. Această analiză evidențiază proprietăți importante ale distorsiunilor, ceea ce deschide posibilități de cercetare viitoare.

- În plus, este propusă o metodă de modelare a formei globale a obiectului transparent, pentru cazurile de utilizare în care nu sunt disponibile cunoștințe prealabile despre obiect (cap. 5). Obiectul este modelat ca un elipsoid, care poate acoperi o mare varietate de

cazuri de utilizare. Folosind o metodologie similară celei utilizate pentru modelul local, se generează un set de date sintetice pentru a evalua metoda.

- Este propusă o metodă de autocalibrare bazată pe învățarea profundă (cap. 6). Modelul de distorsiune ales este mai complex în comparație cu alte metode din literatura de specialitate, deoarece include atât componente globale, cât și locale. Estimarea mai precisă a distorsiunilor este facilitată și de utilizarea sarcinilor auxiliare în arhitectura rețelei.

- Pentru a antrena rețelele neuronale, sunt construite două seturi de date cu imagini distorsionate – cu date simulate și reale – bazate pe măsurători reale ale distorsionării parbrizului.

1.3 Lista de publicații

Contribuțiile acestei teze au fost publicate în lucrările conferinței de mai jos. Lucrările sunt clasificate în funcție de clasamentele *Compute Research and Education (CORE) 2018*[1] (categoriile A*, A, B, C, D). Categoria D include conferințe care nu sunt enumerate în clasamentul CORE.

- Categoria A
  - Szabolcs-Botond Lőrincz, **Szabolcs Pável**, and Lehel Csató. Single View Distortion Correction using Semantic Guidance. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, July 2019.

- Categoria B
  - **Szabolcs Pável**, Csanád Sándor, and Lehel Csató. Distortion Estimation Through Explicit Modeling of the Refractive Surface. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing*, pages 17–28. Springer, September 2019.

- Categoria D
  - **Szabolcs Pável**. An Ellipsoid Object Model of the Refraction Surface. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI)*, volume 2650, pages 272–279. CEUR Workshop Proceedings, January 2020.

[1] CORE Conference Portal: `https://portal.core.edu.au/conf-ranks/`

1.4 Structura tezei

Restul acestei teze este structurat după cum urmează.

În **capitolul 2** stabilim bazele modelelor de camere și ale tehnicilor de calibrare a camerelor. Prezentăm modelele globale clasice utilizate pentru camerele normale și fisheye, care sunt printre cele mai utilizate tipuri de camere. De asemenea, introducem conceptul de distorsiuni ale imaginii și revizuim exemple din literatura de specialitate privind modul în care sunt modelate distorsiunile imaginii. În cele din urmă, ne concentrăm pe modele mai generale de camere, care pot fi aplicate pentru a modela o mare varietate de distorsiuni, inclusiv cele introduse de obiectele transparente din fața camerei. Conținutul acestui capitol se bazează pe (31, 71, 73).

În **capitolul 3** prezentăm concepte fundamentale din învățarea automată și învățarea profundă. Prezentăm funcțiile de bază radiale, pe care le folosim pentru a modela distorsiunile produse de parbrize, atât în metoda de calibrarea inițială, cât și în metoda de calibrare bazată pe învățarea profundă. În continuare, prezentăm elementele de bază ale rețelelor neuronale convoluționale și prezentăm câteva aplicații relevante. Încheiem capitolul cu prezentarea funcțiilor de eroare specifice bazate pe reconstrucția imaginilor, utilizate și de noi pentru formarea rețelelor noastre neuronale. Conținutul acestui capitol se bazează pe (27, 87).

În **capitolul 4** prezentăm o metodă de calibrare inițială care se concentrează pe proprietățile locale ale distorsiunilor. Presupunem că forma globală a obiectului transparent (de exemplu, parbrizul) este cunoscută și modelăm suprafața neuniformă utilizând un model bazat pe funcții de bază radiale. Parametrii modelului sunt optimizați utilizând imagini ale unor ținte de calibrare cu tablă de șah. Testăm metoda noastră pe date sintetice și reale și analizăm distorsiunile imaginii observate. Acest capitol se bazează pe publicația noastră (59).

În **capitolul 5** prezentăm o metodă similară celei din capitolul anterior, dar de data aceasta ne concentrăm pe forma globală a obiectului transparent. Forma globală este aproximată ca un elipsoid, iar noi calibrăm parametrii pe baza țintelor cu tablă de șah. Observăm simetrii în modelul elipsoid și propunem o schemă de regularizare pentru a ghida procesul de optimizare. Acest capitol se bazează pe publicația noastră (58).

În **capitolul 6** ne îndreptăm atenția către autocalibrare și propunem o soluție bazată pe învățarea profundă. În lipsa datelor de instruire disponibile, construim două seturi de date bazate pe măsurători reale ale distorsiunilor din parbrize. Utilizăm din nou funcții de bază radiale pentru a modela distorsiunile neliniare și antrenăm rețelele utilizând funcții de eroare bazate pe reconstrucția imaginilor. Sarcinile auxiliare sunt, de asemenea, integrate în rețea,

unde observăm că atât segmentarea semantică, cât și fluxul optic ne îmbunătățesc rezultatele. Acest capitol se bazează pe publicația noastră (48).

În cele din urmă, **capitolul 7** prezintă concluziile activității noastre.

Conținutul acestei teze se bazează pe o listă de 90 de referințe, dintre care 17 citații reprezintă cele mai noi progrese din domeniu, publicate în ultimii 5 ani, în timp ce altele rezumă lucrările fundamentale din domeniul viziunii computerizate clasice, învățării automate și învățării profunde.

Capitolul 2

Modele de camere

O cameră digitală este un sistem de imagistică care utilizează elemente optice precum lentile și oglinzi pentru a focaliza lumina pe un senzor fotosensibil, mapând lumea 3D pe un plan de imagine 2D. Această mapare duce la pierderea informațiilor despre adâncime, pe care algoritmii de viziune computerizată 3D încearcă să le recupereze folosind mai multe imagini din puncte de vedere diferite. Aceste imagini oferă constrângeri geometrice care ajută la reconstruirea structurii 3D. Un model matematic precis, cunoscut sub numele de *model de cameră*, este esențial pentru descrierea acestui sistem optic și este o componentă cheie a algoritmilor de viziune computerizată 3D.

2.1 Modele clasice de camere

Alegerea modelului de cameră depinde de camera utilizată. *Modelul camerei pinhole* este cel mai utilizat și folosește o proiecție în perspectivă pentru a mapa coordonatele lumii 3D în puncte de imagine 2D. Acest model descrie o cameră ideală și este adesea asociat cu un *model de distorsiune* pentru a corecta deviațiile de la cazul ideal. În cazul sistemelor complexe de camere, cum ar fi camerele fisheye, sunt necesare modele specifice pentru a ține seama de maparea neliniară a pixelilor.

Modelele camerelor sunt parametrice, cu parametri care descriu caracteristicile sistemului camerei, cum ar fi distanța focală, parametrii de distorsiune și poziția camerei în spațiul 3D. *Calibrarea camerei* este procesul de găsire a acestor valori optime ale parametrilor.

2.2 Calibrarea camerei

Metodele de calibrare a camerelor pot fi împărțite în *calibrare inițială* și *autocalibrare*. Calibrarea inițială se realizează în medii controlate, folosind ținte de calibrare precum table de șah, în timp ce autocalibrarea utilizează proprietățile geometrice ale scenei și mișcarea camerei pentru a deduce parametrii de calibrare.

Metoda lui Zhang (88) este o tehnică de calibrare inițială utilizată frecvent, care utilizează ținte planare de tablă de șah. Procesul implică capturarea mai multor imagini ale modelului, extragerea caracteristicilor, calcularea omografiilor și rafinarea parametrilor prin optimizare neliniară pentru a minimiza eroarea de reproiectare.

Metodele de autocalibrare valorifică proprietățile geometrice ale scenei și caută structuri regulate sau se bazează pe mișcarea unei camere. Metoda propusă de Devernay și Faugeras (18), utilizează caracteristici precum liniile drepte din mediu. O altă abordare utilizează potrivirea punctelor din mai multe puncte de vedere pentru a optimiza constrângerile epipolare, așa cum au propus Claus și Fitzgibbon (15).

2.3 Camere Fisheye

Camerele Fisheye au obiective cu unghi larg cu câmpuri de vizualizare de peste 180 de grade. Modelele tradiționale, cum ar fi modelul camerei pinhole, sunt insuficiente din cauza distorsiunilor neliniare semnificative. Pentru camerele fisheye sunt utilizate diferite modele de proiecție, inclusiv proiecții stereografice, echidistante, ortografice și echisolide, fiecare având avantaje specifice.

Kannala et al. (40) au propus un model care utilizează termeni polinomiali pentru a ține seama de distorsiunile radiale mari. Modelele bazate pe funcții raționale, precum modelul de diviziune (division model) al lui Fitzgibbon (22), oferă o formă închisă inversă și sunt utilizate pentru reconstrucția stereo. Modelul Field of View (FOV) al lui Devernay și Faugeras (18) utilizează un singur parametru pentru a descrie obiectivele fisheye și oferă, de asemenea, o formă închisă inversă.

2.4 Modele de camere generalizate

Modelele de camere noncentrale, care iau în considerare deplasarea centrului optic, sunt importante pentru modelarea precisă a obiectivelor cu unghi larg. Gennery (24) a propus un model

pentru obiectivele fisheye cu o funcție de deplasare a pupilei.

Modelele generalizate ale camerelor nu se bazează pe proprietăți fizice, ci tratează sistemul optic ca pe o cutie neagră, oferind flexibilitate cu prețul unei calibrări complexe. Modelul cu două planuri (50) utilizează funcții de interpolare pentru a mapa pixeli 2D în coordonate 3D pe planuri de calibrare, abordând atât distorsiuni globale, cât și locale.

Modelele discrete de camere, cum ar fi modelul *raxel* (28), tratează fiecare pixel individual, încorporând componente geometrice, radiometrice și optice. Aceste modele necesită observații dense pentru calibrare.

2.5 Modelarea distorsiunilor din medii refractive

Camerele care observă scene prin medii refractive, cum ar fi sub apă sau în spatele parbrizelor, se confruntă cu distorsiuni complexe. Acestea pot fi rezolvate prin modelarea explicită a traiectoriei razei refractate sau prin utilizarea unor modele generalizate de camere care tratează mediul refractiv ca parte a sistemului de imagistică.

Agrawal et al. (2) au studiat sisteme cu mai multe straturi de suprafețe refractive plate, descriindu-le drept camere axiale și oferind ecuații analitice de proiecție. Yoon et al. (86) au introdus un model parametric pentru estimarea adâncimii folosind camere stereo în spatele unor obiecte transparente.

Au fost propuse modele generalizate precum modelul în două planuri (80) și modele locale care utilizează B-splines (5) pentru cazuri de utilizare în industria automobilelor. Kim et al. (14, 42) au introdus metode de calibrare a camerelor cu distorsiuni refractive complexe, utilizând funcții de bază radiale.

2.6 Concluzii

Acest capitol a prezentat bazele teoretice ale modelelor și calibrării camerelor. Acesta a acoperit modelul clasic de cameră pinhole, distorsiunile geometrice și tehnicile de calibrare. Pentru sistemele complexe, cum ar fi camerele fisheye, sunt necesare modele specifice pentru a ține seama de distorsiunile mari. Modelele generalizate și noncentrale oferă soluții flexibile pentru diverse sisteme optice. În cele din urmă, au fost discutate metodele de modelare a distorsiunilor din medii refractive, subliniindu-se atât modelarea explicită, cât și modelele generalizate de camere.

Capitolul 3

Învățarea automată și învățarea profundă

Învățarea automată, o ramură a inteligenței artificiale, permite calculatoarelor să învețe din date prin identificarea modelelor statistice. Aceasta presupune formarea de modele pentru sarcini precum clasificarea, regresia și gruparea, utilizând algoritmi de optimizare pentru a minimiza erorile de predicție. Învățarea automată clasică utilizează modele cu un număr limitat de parametri, în timp ce învățarea profundă antrenează rețele neuronale cu milioane de parametri pe seturi mari de date. Acest capitol prezintă conceptele fundamentale ale învățării automate și ale învățării profunde utilizate în activitatea noastră.

3.1 Funcții de bază radiale

Funcțiile de bază radiale (RBF) (3, 11) sunt utilizate pentru interpolarea sau aproximarea datelor dispersate. Având în vedere un set de puncte de date, obiectivul este de a găsi o funcție netedă care să îndeplinească condiția de interpolare în aceste puncte. RBF-urile definesc funcția de interpolare ca fiind suma ponderată a funcțiilor de bază centrate pe punctele de date. Influența unui punct de date asupra valorii interpolate depinde de distanța de la punct. Funcțiile kernel comune includ funcțiile gaussiene, thin plate splines, multiquadrice și multiquadrice inverse.

Pentru interpolare, ponderile sunt setate pentru a satisface condiția de interpolare, care poate fi prezentată sub formă de matrice și rezolvată ca sistem liniar. RBF-urile pot, de asemenea, să aproximeze punctele de date prin minimizarea unei funcții de eroare, care include un termen de eroare a celor mai mici pătrate și un termen de regularizare bazat pe ponderile RBF.

Atunci când este combinată cu o componentă polinomială, funcția de interpolare are putere de reprezentare globală. În acest caz, funcția de eroare include o constrângere de ortogonalitate pentru a asigura că proprietățile globale sunt caracterizate de componenta polinomială.

3.2 Optimizarea bazată pe gradient a modelelor parametrice

Tehnicile de optimizare sunt esențiale în învățarea automată pentru găsirea parametrilor corecți ai modelului prin minimizarea unei funcții de eroare. În învățarea profundă, antrenarea implică actualizarea parametrilor rețelei neuronale utilizând gradientul unei funcții de eroare specifice sarcinii, de obicei cu ajutorul algoritmului *Stochastic Gradient Descent* (SGD). Metodele de optimizare de ordinul întâi, precum SGD, utilizează prima derivată a funcției de eroare și sunt adecvate pentru învățarea profundă datorită numărului mare de parametri ai modelului. Variante ale SGD, cum ar fi SGD cu momentum și optimizatorul Adam (43), abordează probleme legate de setările inițiale ale ratei de învățare și oferă rate de învățare adaptive.

Tehnicile de optimizare de ordinul doi iau în considerare curbura funcției obiectiv prin intermediul matricei Hessian, oferind o convergență mai rapidă pentru problemele cu mai puțini parametri. Metoda Newton-Raphson și forma sa simplificată pentru probleme de pătratele minime neliniare, algoritmul Gauss-Newton, sunt frecvent utilizate. Metodele quasi-Newton, cum ar fi algoritmul Broyden-Fletcher-Goldfarb-Shanno (BFGS) și varianta sa cu memorie limitată LBFGS, aproximează matricea Hessiană și sunt adecvate pentru problemele în care calcularea Hessianului complet nu este practică.

3.3 Arhitecturi convoluționale

Rețelele neuronale convoluționale (CNN) sunt arhitecturi de ultimă generație pentru sarcinile de vedere pe calculator. Acestea prelucrează imaginile prin straturi de convoluții, reducere sau creștere a rezoluției și funcții de activare. CNN-urile sunt eficiente datorită capacității lor de a modela interacțiunile locale și invarianța la translație. Principalele straturi includ straturi convoluționale, funcții de activare, straturi de normalizare, straturi de pooling, straturi de upsampling și straturi complet conectate.

Rețelele reziduale (ResNets) (33, 34), cunoscute pentru blocurile lor reziduale și conexiunile de tip skip, abordează problema gradientului evanescent, permițând antrenarea eficientă a rețelelor mai profunde. Diferitele modele ResNet, precum ResNet-18, ResNet-34, ResNet-50,

ResNet-101 și ResNet-152, variază în profunzime și complexitate, dar au o structură comună, cu un strat convoluțional inițial urmat de etape de blocuri reziduale.

3.4 Sarcini de viziune artificială

Rețelele neuronale au fost aplicate cu succes la diverse sarcini de viziune computerizată, inclusiv segmentarea semantică, estimarea fluxului optic și calibrarea camerei. Segmentarea semantică clasifică fiecare pixel dintr-o imagine în clase semantice, esențiale pentru sistemele de percepție auto. Arhitecturi precum *U-Net* (63) și *DeepLabv3* (13) utilizează structuri codificator-decodificator și, respectiv, convoluții dilatate pentru a menține detaliile de înaltă rezoluție, capturând în același timp contextul global.

Fluxul optic descrie mișcarea pixelilor între imagini și poate fi dispersat sau dens. *FlowNet* (20) și succesorul său *FlowNet 2.0* (37) au fost printre primele abordări de învățare profundă pentru estimarea fluxului optic dens, utilizând arhitecturi codificator-decodificator și procesare în cascadă pentru a gestiona deplasări mari și mici.

Metodele de învățare profundă pentru calibrarea camerelor pot estima parametrii intrinseci și extrinseci din imagini unice. Primele metode, precum *DeepFocal* (82), au presupus un model simplificat de cameră pinhole și au prezis un set limitat de parametri. Abordările ulterioare, cum ar fi *FishEyeRectNet* (85), au introdus diverse îmbunătățiri, inclusiv utilizarea de sarcini auxiliare și funcții de eroare bazate pe reconstrucția imaginilor.

3.5 Instruire bazată pe reconstrucția imaginilor

Atunci când supravegherea directă a parametrilor de calibrare a camerei este dificilă, pot fi utilizate funcții de eroare bazate pe reconstrucția imaginii. *Rețelele de transformare spațială* (Spatial Transformer Networks, STN) (39) permit deformarea diferențiabilă a imaginii, ceea ce face posibilă antrenarea rețelelor de la un capăt la altul folosind pierderi de reconstrucție. STN-urile constau dintr-o rețea de localizare, un generator de grilă și un eșantionator diferențiabil.

În formarea autosupravegheată, cum ar fi *SfmLearner* (90), STN-urile sunt utilizate pentru a deforma imaginile sursă pentru a se potrivi cu vederile țintă, utilizând adâncimea prezisă și mișcarea camerei. Eroarea de reconstrucție se bazează pe diferența dintre vederile țintă reale și cele generate. Combinarea pierderii L1 cu Structural Similarity Index Measure (SSIM) sau cu SSIM pe mai multe scări (MS-SSIM) (81) oferă un semnal de învățare mai robust prin luarea în considerare a percepției umane a calității imaginii.

3.6 Concluzii

Acest capitol a acoperit conceptele esențiale de învățare automată și învățare profundă relevante pentru activitatea noastră. Funcțiile de bază radiale (RBF) sunt utile pentru interpolarea funcțiilor și pentru aproximarea și tehnicile de optimizare bazate pe gradient sunt esențiale pentru formarea modelelor de învățare automată. Rețelele neuronale convoluționale (CNN) rămân arhitectura dominantă pentru sarcinile de viziune computerizată. Rețelele neuronale sunt eficiente pentru autocalibrare și pot fi antrenate utilizând funcții de eroare bazate pe reconstrucția imaginii cu ajutorul rețelelor de transformatoare spațiale.

Capitolul 4

Modelarea explicită a suprafeței de refracție

Acest capitol prezintă publicația noastră intitulată *Distortion Estimation Through Explicit Modeling of the Refractive Surface* (59).

Această lucrare se concentrează în principal pe sistemele de camere utilizate în aplicații auto, unde camerele sunt montate în spatele parbrizului sau al altor capace de protecție. Prezența acestor materiale refractive complică modelarea geometrică a sistemului de camere datorită refracției luminii, care duce la distorsiuni ale imaginii atunci când lumina intră sau iese dintr-un mediu mai dens, provocând schimbări direcționale.

Atunci când lumina trece printr-un material refractiv, aceasta își schimbă direcția, rezultând distorsiuni ale imaginii. Acest fenomen îngreunează utilizarea modelelor globale, centrale de cameră, deoarece refracțiile fac ca sistemul optic să fie dificil de caracterizat. Literatura de specialitate abordează de obicei această problemă folosind una dintre cele două abordări: fie prin modelarea explicită a refracțiilor, fie prin aplicarea unor modele de cameră generalizate. Pentru aplicațiile auto, forma globală a parbrizului este în general cunoscută prin modele de proiectare asistată de calculator (CAD), deși pot exista neregularități locale. Presupunând că forma globală este cunoscută, propunem un model de cameră care ia în considerare în mod explicit refracțiile luminii și include un model local pentru tratarea iregularităților de suprafață. Această metodă poate fi utilizată pentru calibrarea inițială a camerelor din spatele obiectelor transparente.

Construim *modelul forward* $f_\theta(\boldsymbol{p}) : \Omega \to \mathbb{R}^3$, care mapează un pixel din imagine către un punct din scenă, luând în considerare parametrii camerei, mediul de refracție și caracteristicile

scenei, denumite colectiv θ. Această funcție este implementată ca un algoritm de raycasting, permițând generarea de imagini având în vedere un set de parametri. Prin inversarea modelului, ajustăm parametrii mediului refractiv la punctele deplasate observate. Construim o rețea RBF (Radial Basis Function) (6) al grosimii mediului de refracție și folosim estimarea de verosimilitate maximă (ML) pentru a deduce parametrii optimi care au cauzat distorsiunile.

Contribuțiile acestui capitol sunt:

- Introducerea unui *model local* pentru suprafețele refractive inegale, în contrast cu modelele globale predominante în literatura de specialitate.
- Ne concentrăm asupra scenariilor în care forma globală a obiectului transparent este cunoscută, prezentând un *model explicit de refracție*.
- Demonstrarea faptului că astfel de modele reduc semnificativ eroarea de calibrare.
- O *analiză* a distorsiunilor parbrizului, care oferă o perspectivă asupra acestor sisteme optice.

4.1 Model de suprafață refractivă

Mediul refractiv este modelat ca o felie de con gros, care se bazează pe forma reală a obiectului de sticlă utilizat în experimentele noastre. Conurile interior și exterior au aceeași deschidere, iar centrele sunt astfel încât grosimea mediului este constantă. Conul este poziționat cu axa sa principală paralelă cu axa y a camerei și se micșorează în direcția y pozitivă.

Pentru a ține seama de suprafața neuniformă, se adaugă o suprafață parametrică în direcția radială la conului. Acest decalaj radial este definit cu ajutorul unei rețele RBF (Radial Basis Function), care este aleasă pentru capacitățile sale universale de aproximare a funcțiilor (6). Centrele RBF sunt plasate pe o grilă regulată peste regiunea de intrare, iar ponderile $w_{ij} \in \mathbb{R}$ sunt reglate pentru a modela suprafețe complexe. Deplasarea radială la un anumit punct de con este ieșirea rețelei RBF cu nuclee gaussiene:

$$\Phi(\boldsymbol{s}') = \sum_{i,j=1}^{N} w_{ij}\phi(\|\boldsymbol{s}_{ij} - \boldsymbol{s}'\|), \quad \text{unde } \phi(r) = \exp\left(-\frac{r^2}{2\beta}\right) \tag{4.1}$$

Pentru a calcula coordonatele carteziene ale unui punct de pe con, parametrizat de o înălțime s_1 și un unghi s_2, se calculează mai întâi decalajul RBF și apoi se adaugă la rază. Normalele de suprafață ale conului exterior sunt calculate ca produs vectorial al derivatelor parțiale ale

coordonatelor carteziene în raport cu parametrii. Acest produs vectorial depinde de ponderile RBF, care sunt utilizate ca parametri ai modelului în timpul minimizării. Modificarea ponderilor RBF modifică normalele suprafeței, schimbând astfel direcția razelor de lumină refractate și, în consecință, vectorii de distorsiune.

4.2 Modelul Raycasting

Modelul raycasting descrie procesul de asociere a unui pixel din imagine cu un punct 3D de pe un obiect, cum ar fi un o țintă cu tablă de șah. Într-o configurație fără distorsiuni, acest lucru poate fi realizat utilizând modelul camerei pinhole, care utilizează o proiecție în perspectivă și un set de operații liniare pentru a descrie această relație. Cu toate acestea, în prezența unei suprafețe refractive, această descriere geometrică simplă nu este valabilă și sunt necesare etape suplimentare.

Fiecare rază începe din centrul camerei, presupunând că parametrii intrinseci ai camerei – distanțele focale și punctul principal – sunt cunoscuți. Toate punctele 3D sunt exprimate în sistemul de coordonate al camerei. Folosind matricea intrinsecă al camerei, orice coordonată $\boldsymbol{p}$ a pixelului poate fi convertită în coordonate metrice, care, după normalizare, corespund vectorului de direcție $\boldsymbol{r_{cam}}$ al razei de lumină care trece prin pixelul selectat.

Raza de lumină atinge mai întâi partea interioară a suprafeței de refracție, intersectând conul la $\boldsymbol{x}_i$ și având normala $\boldsymbol{n}_i$. Direcția $\boldsymbol{r}_m$ a razei de lumină refractate în interiorul mediului este calculată folosind legea lui Snell. Cunoscând geometria corpului de refracție și noua direcție a razei refractate, se identifică locul $\boldsymbol{x}_o$ în care raza atinge suprafața exterioară și se calculează a doua refracție. Se determină apoi direcția de ieșire a razei de lumină, $\boldsymbol{r}_o$. Această a doua refracție este modulată de direcția normalei $\boldsymbol{n}_o$, parametrizată de rețeaua RBF.

În cele din urmă, raza de lumină care iese intersectează ținta de calibrare, care este definită printr-o rotație 3D și o translație a centrului plăcii în raport cu sistemul de coordonate al camerei. Punctul de intersecție $\boldsymbol{x}_t$ este calculat ca intersecție a unei linii și a unui plan. Pentru o manipulare mai ușoară, se definește un sistem local de coordonate 2D pe planul obiectului, cu originea în centrul plăcii și cu axele corespunzătoare direcțiilor orizontală și verticală ale grilei tablei de șah. Coordonatele locale ale unui punct 3D $\boldsymbol{x}_t$ sunt notate $\boldsymbol{x}_{cb}$.

4.3 Optimizarea parametrilor de suprafață

Estimarea distorsiunilor imaginii implică găsirea parametrilor de suprafață care au generat un set de imagini de calibrare, folosind un model de tablă de șah ca țintă de calibrare. Minimizarea Gradient Descent este utilizată pentru a determina ponderile RBF optime, în timp ce alți parametri, inclusiv caracteristicile intrinseci ale camerei, dimensiunile conurilor interioare și exterioare, precum și poziția și dimensiunea țintei de calibrare, sunt presupuse a fi cunoscute.

Pentru fiecare imagine de calibrare, sunt identificate coordonatele pixelilor și ordinea colțurilor modelului de tablă de șah. Coordonatele locale corespunzătoare ale colțurilor detectate pe planul obiectului sunt date de distanța lor față de centrul plăcii. Funcția de raycasting, parametrizată de ponderile RBF, mapează un pixel de intrare la coordonatele locale ale punctului mondial corespunzător pe modelul de tablă de șah țintă. Funcția de eroare este definită ca distanța $\mathcal{L}_2$ între coordonatele locale estimate ale unui colț și coordonatele locale reale.

4.4 Rezultatele calibrării modelului suprafeței refractive

Deoarece nu există niciun set de date publice de referință pentru problema noastră, am creat propria noastră configurație experimentală. Evaluăm algoritmul pe două seturi de date: un set de date sintetice fără zgomot și o configurație experimentală reală. În cazul sintetic, arătăm că algoritmul nostru este capabil să găsească parametrii optimi care au generat o imagine dată, chiar și cu neregularități mari pe suprafața exterioară. În al doilea caz, prezentăm o configurație experimentală și demonstrăm că algoritmul poate reduce erorile de reconstrucție în scenarii reale. În setul de date sintetice, aplicăm modelul de generare avansată a imaginii pentru a reda imagini sintetice. Parametrii camerei, ai suprafeței refractive și ai modelului de tablă de șah sunt setați la valori similare cu cele din experimentul din lumea reală. Folosind o grilă RBF de dimensiune 4×4, eșantionăm ponderile dintr-o distribuție gaussiană și generăm imagini sintetice folosind poziții aleatorii pentru ținta de calibrare. Optimizarea este efectuată, iar eroarea finală este stocată ca eroare medie pătratică (RMSE) între colțurile previzionate și cele reale ale tabloului de șah.

În setul de date reale, un modul de cameră Raspberry Pi Camera Module v2 captează imaginile de tip tablă de șah. Camera are un senzor de 3.68×2.76 mm și înregistrează imagini la o rezoluție de 3280×2464 pixeli. Camera este calibrată folosind metoda lui Zhang (88), rezultând o distanță focală de 2558.36 pixeli și un punct principal la (1666.03, 1273.65). După

calibrare, un obiect de sticlă în formă de con este plasat în fața camerei. Se captează imagini, iar procesul de optimizare se efectuează pe diferite seturi de imagini.

Rezultatele arată o reducere semnificativă a erorilor la pixelii reproiectați, demonstrând eficacitatea modelului. Erorile sunt evaluate în termeni de RMSE între pozițiile observate și cele prezise ale colțurilor tablei de șah. Eroarea inițială, luând în considerare modelul fără distorsiuni, este comparată cu eroarea finală obținută utilizând modelul de suprafață Cone + RBF optimizat. Metoda îmbunătățește distanțele 3D, rezultând un model generativ mai precis al procesului de formare a imaginii. Analiza distorsiunilor arată o dependență liniară de adâncimea inversă, sugerând un potențial pentru modele de camere mai simple.

Majoritatea metodelor de estimare a distorsiunilor modelează în mod direct deplasările pixelilor pe planul imaginii, definind o hartă de distorsiune unică și fixă pentru o anumită cameră. În schimb, modelul nostru estimează harta de distorsiune prin modelarea explicită a materialului refractiv utilizând un algoritm de raycasting. Această abordare oferă un model generativ unificat și coerent pentru distorsiuni. Modelul fizic introduce o componentă dependentă de adâncime în harta de distorsiune, necesitând distanța unui punct 3D pentru a determina distorsiunea imaginii.

Pentru a calcula vectorul de distorsiune a imaginii, pornim de la un pixel distorsionat al imaginii. Folosind pixelul distorsionat $\boldsymbol{p}_d$, algoritmul de raycasting calculează coordonatele 3D $\boldsymbol{x}_t$ ale punctului obiect la o anumită distanță. Coordonatele pixelului nedistorsionat $\boldsymbol{p}_u$ sunt apoi calculate folosind modelul camerei pinhole. Vectorul de distorsiune $\Delta\boldsymbol{p}$ este dat de diferența dintre coordonatele distorsionate și cele nedistorsionate.

Câmpul vectorial al distorsiunilor pentru setul de date real este vizualizat, arătând dependența distorsiunilor de adâncimea pixelilor. Pixelii din mijloc prezintă o distorsiune redusă din cauza razelor aproape ortogonale față de suprafața de refracție, în timp ce pixelii laterali prezintă distorsiuni mai mari din cauza refracțiilor semnificative. Dependența liniară a distorsiunii de adâncimea inversă este confirmată, indicând faptul că punctele de obiect mai apropiate provoacă distorsiuni mai mari.

4.5 Concluzii

În acest capitol, am prezentat un model de cameră care ia în considerare suprafața neuniformă a unui obiect transparent utilizând o rețea RBF. Prin modelarea explicită a refracțiilor luminii la suprafața obiectului și prin utilizarea legii lui Snell, am dezvoltat un model cuprinzător care ia

în considerare atât forma globală, cât și proprietățile locale ale obiectului. Parametrii modelului au fost estimați cu ajutorul imaginilor de calibrare cu modele de tablă de șah.

Experimentele noastre au demonstrat că modelul propus reduce semnificativ eroarea la pixelii reproiectați. Modelul atenuează în mod eficient distorsiunea introdusă de curbura orizontală a obiectului în formă de con, rezultând erori care seamănă cu o distribuție gaussiană izotropică. Dependența liniară a distorsiunilor de adâncimea inversă oferă oportunități de dezvoltare a unor modele de cameră mai simple, care pot fi integrate mai ușor în sistemele de viziune computerizată.

O limitare a metodei noastre este forma fixă a bazei, care în acest studiu a fost un con. Deși această formă se apropie foarte mult de obiectul real utilizat în experimente, parametrizarea formei bazei ar putea extinde aplicabilitatea modelului la alte geometrii. Cu toate acestea, creșterea numărului de parametri liberi introduce, de asemenea, dificultăți suplimentare de modelare, făcând optimizarea mai dificilă. În rezumat, această lucrare oferă o explorare detaliată a modelării explicite a suprafețelor refractive, contribuind la domeniul calibrării camerelor pentru sistemele care operează în spatele materialelor transparente. Abordarea are potențialul de a îmbunătăți acuratețea sistemelor de vizionare în industria auto și în alte aplicații în care distorsiunile de refracție sunt predominante.

Capitolul 5

Model de obiect elipsoid

Acest capitol prezintă publicația noastră intitulată *An Ellipsoid Object Model of the Refraction Surface* (58).

Distorsiunile geometrice pot apărea atunci când camera este plasată în spatele unui element de protecție, cum ar fi parbrizul unei mașini. Aceste distorsiuni sunt influențate de proprietățile globale ale obiectului (de exemplu, poziția față de cameră, curbura suprafeței și grosimea materialului), precum și de neregularitățile suprafeței, rezultând distorsiuni locale. În cap. 4, am modelat aceste neregularități utilizând o rețea RBF, presupunând în același timp că proprietățile globale ale obiectului refractiv sunt cunoscute. Această presupunere limitează aplicabilitatea modelelor de cameră.

În această lucrare, abordăm proprietățile globale ale obiectului refractiv folosind o metodologie similară celei din cap. 4. Modelăm suprafața mediului refractiv ca un *elipsoid*, care poate aproxima o varietate de obiecte în vederea camerei. Modelul este conceput pentru a putea fi compus cu modelul de rețea RBF al suprafeței locale.

Contribuțiile acestui capitol sunt:

- Este propus un *model* global bazat pe o *formă elipsoidală* și este definit algoritmul de raycasting.
- Capitolul abordează problema apariției *simetriilor* în procesul de estimare a distorsiunilor, unde este inclus un termen de regularizare pentru a ghida minimizarea.
- Metoda este *evaluată* pe un set de date sintetice.

5.1 Modelul elipsoid

Modelăm obiectul refractiv ca fiind spațiul dintre două elipsoide care au aceeași poziție centrală și aceeași orientare. Elipsoidul interior are semi-axele a, b, c, iar semi-axele elipsoidului exterior sunt definite prin adăugarea unei grosimi mici t la fiecare semi-axă a elipsoidului interior. Parametrii modelului de obiect elipsoid sunt poziția centrului, orientarea, semi-axele și grosimea.

Pentru a simplifica calculele ulterioare, elipsoidul este văzut ca o imagine afină a unei sfere unitare centrate la origine. Transformarea implică o matrice 3×3 și un vector de translație. Folosind această mapare afină, operațiile necesare, inclusiv intersecția cu o rază și evaluările normale ale suprafeței, pot fi reduse la operații pe sfera unitară.

Scopul modelului camerei este de a asocia pixelii cu razele de lumină din lumea exterioară. Într-o configurație fără distorsiuni, raza de lumină din centrul camerei trece prin pixelul imaginii, conform modelului camerei pinhole. În modelul nostru, direcția razei de lumină originale se schimbă atunci când intră sau iese din obiectul refractiv, calculată folosind legea refracției a lui Snell. Această schimbare este o funcție a razei incidente, a normalei suprafeței la punctul de intersecție și a indicelui de refracție relativ al materialelor. Procesul de raycasting este complet diferențiabil, permițând optimizarea pe bază de gradient a parametrilor modelului elipsoid. Metoda este implementată în PyTorch pentru a utiliza diferențierea automată pentru optimizare.

5.2 Simetrii ale modelului de obiect elipsoid

Modelul elipsoid supra-parametrizează distorsiunile imaginii, rezultând simetrii în modelul fizic. Distorsiunile observate sunt invariante la anumite transformări ale obiectului. În consecință, setul complet de parametri nu poate fi recuperat fără cunoștințe prealabile despre obiect. Deși estimarea distorsiunilor poate fi suficientă în unele cazuri, reconstrucția unui model 3D apropiat al obiectului poate fi de dorit. Identificarea și tratarea acestor simetrii cu ajutorul tehnicilor de regularizare este esențială.

Un exemplu intuitiv al acestor simetrii este un caz 2D în care obiectul refractiv este un cerc gros. Luăm în considerare două variabile: distanța relativă a centrului cercului față de centrul camerei și raza cercului. Această analiză se aplică, de asemenea, poziției centrale a elipsoidului și lungimii semi-axelor în 3D. Prin compararea erorii de distorsiune în raport cu o configurație de referință pentru diferiți parametri, observăm că se pot obține erori de distorsiune reduse prin ajustarea corespunzătoare a ambilor parametri.

5.3 Optimizarea parametrilor modelului

Utilizăm o configurație standard de calibrare a camerei statice pentru a estima parametrii modelului. Un model de tablă de șah plană servește ca obiect țintă, având dimensiuni cunoscute ale pătratelor. Folosind inversarea modelului pe baza imaginilor modelelor de tablă de șah, parametrii modelului elipsoid pot fi recuperați prin minimizarea bazată pe gradient.

Funcția de eroare cuprinde un termen de reconstrucție și un termen de regularizare. Eroarea de reconstrucție este eroarea medie pătratică dintre coordonatele estimate și cele reale ale colțului plăcii de șah. Termenul de regularizare include cunoștințe anterioare ca o constrângere asupra distanței dintre centrul camerei și punctul în care axa principală intersectează suprafața interioară a elipsoidului. Funcția de eroare este minimizată utilizând metoda de optimizare L-BFGS, aleasă pentru eficiența și compatibilitatea sa cu diferențierea automată.

5.4 Concluzii

Acest capitol a abordat scenariul în care o cameră este plasată în spatele unui obiect transparent cu o formă globală necunoscută. Am modelat obiectul ca un elipsoid și am utilizat tehnici de învățare automată pentru a estima parametrii modelului. Modelul este compatibil cu modelul de suprafață RBF din capitolul 4, permițând modelarea simultană a proprietăților globale și locale ale suprafeței refractive. Am analizat potențialele cazuri de eșec, am propus o regularizare pentru a le aborda și am testat metoda pe un set de date sintetice. Metoda a obținut o aproximare apropiată a suprafeței obiectului în vederea camerei.

Capitolul 6

Estimarea distorsiunii dintr-o singură imagine

Acest capitol prezintă publicația *Single View Distortion Correction using Semantic Guidance* (48).

Capitolele anterioare au discutat despre o metodă de *calibrare inițială (offline)* care abordează distorsiunile cauzate de obiecte transparente, cum ar fi parbrizele mașinilor. Modelul camerei a fost împărțit în distorsiuni locale și o optimizare globală a formei, după cum se detaliază în capitolele 4 și, respectiv, 5. În cazul automobilelor, camerele trebuie să își mențină funcționalitatea în timp, în ciuda vibrațiilor și a schimbărilor de temperatură, ceea ce necesită o componentă de calibrare online bazată pe *autocalibrare*, care funcționează fără obiective specifice de calibrare.

Propunem o metodă de autocalibrare folosind *învățarea profundă* cu un model de distorsiune bazat pe *thin plate spline* (TPS). Rețeaua neuronală prezice parametrii modelului de distorsiune, inclusiv punctele de control pentru componentele locale și coeficienții polinomiali pentru componentele globale. Experimentele demonstrează capacitatea modelului de a estima distorsiuni complexe, ceea ce îl face potrivit pentru aplicații practice în sistemele de conducere autonomă și în alte domenii în care sistemele de camere sunt expuse la condiții variabile pe perioade lungi.

Contribuțiile prezentate în acest capitol sunt:

- Capitolul prezintă o *abordare scalabilă de învățare profundă* care poate corecta distorsiunile. În timp ce metodele de învățare profundă prezentate în literatura de specialitate

prezic, de obicei, doar un număr mic de parametri pentru un model global de cameră, modelul de distorsiune propus este aplicabil distorsiunilor complexe (inclusiv cele locale).

- Similar altor metode din literatura de specialitate, sunt construite *două seturi de date* care cuprind imagini din lumea reală (KITTI Odometry (23)) și sintetizate (Carla (21)) și segmentarea semantică corespunzătoare, pe care sunt aplicate distorsiuni parametrice eșantionate dintr-o distribuție derivată din măsurători din lumea reală în prezența diferitelor parbrize.

- Rețelele sunt antrenate într-o manieră end-to-end fără a utiliza distorsiuni greu de obținut ca supraveghere și, în schimb, progresele recente în eșantionarea imaginilor diferențiabile sunt valorificate pentru a formula o funcție de eroare bazată pe *reconstrucția imaginii*.

- Rezultatele arată că *sarcinile auxiliare* (segmentarea semantică și fluxul optic) îmbunătățesc calitatea predicțiilor.

6.1 Set de date cu distorsiuni ale parbrizului

Pentru a valida capacitatea modelului nostru de a corecta distorsiunile din imagini, am construit două seturi de date, *Distorted Carla (DC)* și *Distorted KITTI (DK)*, conform metodologiilor stabilite. Aceste seturi de date au fost concepute pentru a testa performanța modelului atât în scenarii sintetice, cât și în lumea reală, oferind o evaluare cuprinzătoare a capacităților sale.

Folosim un set de date proprietar de la 240 de parbrize auto, în care imaginile au fost capturate cu și fără parbriz și distorsiunea la nivel de pixel este măsurată. Am potrivit o funcție polinomială de ordin înalt la aceste măsurători, iar noi distorsiuni au fost generate prin eșantionarea și perturbarea coeficienților polinomiali. Această metodă a menținut o variabilitate realistă între imagini, asigurând că distorsiunile sintetice se aseamănă foarte mult cu cele întâlnite în condiții reale.

Distorted Carla (DC) cuprinde 10000 de imagini sintetice și etichete semantice, generate utilizând simulatorul Carla (21). Imaginile, capturate la 5 cadre pe secundă într-un mediu prestabilit, au fost împărțite în 8000 de probe de antrenare și 2000 de probe de validare. Setul de date a inclus imagini RGB și hărți de segmentare semantică, care au fost utilizate pentru a oferi un context suplimentar pentru modelul de corectare a distorsiunilor. Flexibilitatea simulatorului Carla ne-a permis să creăm un set divers de condiții, inclusiv diferite momente ale zilei, condiții meteorologice și elemente dinamice precum vehicule și pietoni.

Pentru datele din lumea reală, am utilizat setul de date KITTI Odometry (23). Acest set de date include secvențe capturate de la un vehicul în mișcare în diferite medii urbane. Am redus rezoluția imaginilor și am aplicat distorsiuni sintetice, creând setul de date *Distorted KITTI (DK)* cu 10684 imagini de antrenare și 4539 imagini de validare.

Fluxul optic a fost încorporat ca o sarcină auxiliară, folosind triplete de imagini ca intrare pentru o rețea de estimare a fluxului. Prin includerea fluxului optic, am furnizat modelului informații temporale, care sunt deosebit de utile pentru înțelegerea scenelor dinamice și îmbunătățirea preciziei corectării distorsiunilor.

6.2 Model de distorsiune

Bookstein (7) a arătat că o pereche de *thin plate spline* (TPS) poate modela deformările 2D. Am modelat distorsiunile geometrice utilizând perechi TPS, care sunt deosebit de eficiente pentru reprezentarea deformărilor netede și continue. Această alegere a modelului ne permite să gestionăm atât distorsiunile globale care afectează întreaga imagine, cât și distorsiunile locale care sunt limitate la anumite regiuni.

Coordonatele transformate $\boldsymbol{f}_{tps}(\boldsymbol{G}_i)$ la coordonata imaginii $\boldsymbol{G}_i = [x_i, y_i]^\top$ presupunând n puncte de control sunt definite în felul următor:

$$\boldsymbol{f}_{tps}(\boldsymbol{G}_i) = \boldsymbol{A} \begin{bmatrix} \boldsymbol{G}_i \\ 1 \end{bmatrix} + \sum_{k=1}^{n} \phi(\|\boldsymbol{p}'_k - \boldsymbol{G}_i\|) \cdot \boldsymbol{w}_k, \quad \text{unde } \phi(r) = r^2 \log r \tag{6.1}$$

Am folosit 16 puncte de control, distribuite uniform pe o grilă 4×4. Transformarea afină $\boldsymbol{A}$ a modelat distorsiunile globale, în timp ce nucleul de bază radial $\phi(r)$ și matricea coeficienților de deformare $\boldsymbol{W}$ au capturat deformările locale. Această combinație de componente globale și locale permite modelului nostru să gestioneze o gamă largă de tipuri de distorsiuni, de la transformări liniare simple la deformări neliniare complexe.

Flexibilitatea modelului TPS îl face ideal pentru aplicații în care distorsiunile nu sunt uniforme pe întreaga imagine. De exemplu, în cazul automobilelor, distorsiunile cauzate de un parbriz pot varia semnificativ în funcție de forma și poziția acestuia față de cameră. Utilizând TPS, putem modela cu precizie aceste variații și le putem corecta într-o manieră uniformă.

6.3 Arhitectura propusă

Arhitectura noastră end-to-end ia ca input o singură imagine distorsionată $\boldsymbol{I}$ și emite imaginea nedistorsionată $\boldsymbol{I}'$ și, opțional, etichetele sale semantice. Aceasta urmează o structură codificator-decodificator cu sarcini auxiliare care oferă un context suplimentar pentru procesul de corectare a distorsiunilor.

O rețea ResNet-18 (33) preantrenată pe ImageNet (65) a servit drept rețea principală. Această rețea extrage caracteristici de nivel scăzut din imaginea de intrare, care sunt apoi utilizate de decodor pentru a estima și corecta distorsiunile. Atunci când se utilizează fluxul optic, rețeaua a procesat triplete de imagini concatenate. Această modificare a permis rețelei centrale să gestioneze mai multe cadre simultan, oferind un set mai bogat de caracteristici pentru corectarea distorsiunilor.

Rețeaua de segmentare semantică a mărit hărțile de caracteristici și le-a concatenat cu imaginea de intrare pentru predicția distorsiunilor. Prin includerea segmentării semantice, modelul a putut valorifica informații de nivel înalt despre scenă, cum ar fi locația obiectelor și a limitelor. Aceste informații ajută modelul să ia decizii mai informate cu privire la modul de corectare a distorsiunilor, în special în zonele cu structuri complexe.

Rețeaua de corectare a distorsiunilor a urmat arhitectura Spatial Transformer Network (39). Aceasta a localizat punctele de control, a calculat grila de eșantionare și a eșantionat imaginea distorsionată pentru a crea imaginea corectată. Natura diferențiabilă a rețelei de transformare spațială permite întregului proces să fie antrenat end-to-end, asigurându-se că toate componentele funcționează împreună fără probleme.

Încorporarea unor sarcini auxiliare precum segmentarea semantică și fluxul optic nu numai că îmbunătățește performanța modelului, dar oferă și rezultate suplimentare care pot fi utile în alte aplicații. De exemplu, hărțile de segmentare semantică pot fi utilizate pentru detectarea obiectelor și înțelegerea scenei, în timp ce fluxul optic poate furniza informații despre mișcăriile în scenă.

6.4 Antrenarea modelului

Parametrii modelului au fost inițializați utilizând inițializarea uniformă "He" (32). Am utilizat diferite funcții de eroare, inclusiv o eroare de reconstrucție a imaginii bazată pe MS-SSIM (81) și o eroare de grilă formulată ca eroarea pătratică medie între grilele de eșantionare estimate și cele ale valorilor de referință. Aceste funcții de eroare asigură faptul că modelul învață să

producă imagini precise nedistorsionate și să alinieze grila de eșantionare cât mai aproape de valorile de referință.

Funcția finală de eroare a fost o sumă ponderată a pierderilor de reconstrucție a imaginii, de grilă și de segmentare semantică, optimizată utilizând Adam (43) cu un batch size de 8 și rate de învățare specifice pentru diferite componente ale rețelei. Această combinație de funcții de eroare echilibrează nevoia de reconstrucție precisă a imaginii cu cerința de a alinia grila de eșantionare și de a produce etichete semantice corecte.

În timpul antrenării, am experimentat cu diferite setări pentru a găsi configurația optimă. Am constatat că utilizarea atât a funcției de eroare de reconstrucție, cât și a celei de grilă a oferit cele mai bune rezultate, deoarece fiecare funcție de eroare o completează pe cealaltă. Funcția de eroare de reconstrucție asigură păstrarea calității generale a imaginii, în timp ce funcția de eroare de grilă se concentrează pe alinierea precisă a distorsiunilor.

Includerea segmentării semantice și a fluxului optic ca sarcini auxiliare a îmbunătățit și mai mult performanța modelului. Aceste sarcini furnizează informații suplimentare care ajută modelul să înțeleagă mai bine scena, conducând la o corecție mai precisă a distorsiunilor. Prin instruirea modelului de la un capăt la altul, ne-am asigurat că toate componentele funcționează perfect împreună, rezultând un sistem robust și fiabil de corectare a distorsiunilor.

6.5 Rezultatele antrenării

Am antrenat rețelele pe setul Distorted Carla Train și le-am testat pe seturile Distorted Carla Test și Distorted KITTI Test. Reglarea fină pe setul Distorted KITTI Train a îmbunătățit și mai mult rezultatele, demonstrând capacitatea modelului de a se adapta la diferite seturi de date.

Evaluarea cantitativă a utilizat norma de distorsiune reziduală, o metrică care măsoară distanța medie dintre punctele distorsionate și nedistorsionate ale grilei. Metoda noastră a redus în mod semnificativ distorsiunile pe ambele seturi de date, cea mai bună performanță fiind obținută utilizând o combinație a funcțiilor de eroare de reconstrucție și de grilă cu fluxul optic ca sarcină auxiliară.

Fără reglare fină, rețeaua antrenată pe Distorted Carla s-a transferat bine la Distorted KITTI, cu excepția cazului în care fluxul optic a fost utilizat ca sarcină auxiliară. Reglarea fină a îmbunătățit rezultatele în toate configurațiile, indicând faptul că modelul se poate adapta la diferite tipuri de date.

Utilizarea sarcinilor auxiliare a îmbunătățit performanța, fluxul optic producând cele mai bune rezultate. Această îmbunătățire se datorează probabil informațiilor temporale suplimentare furnizate de fluxul optic, care ajută modelul să înțeleagă mișcarea și dinamica scenei. Segmentarea semantică a îmbunătățit, de asemenea, performanța, furnizând informații contextuale de nivel înalt care ajută modelul să ia decizii mai informate cu privire la modul de corectare a distorsiunilor.

Atât reconstrucția, cât și pierderile de grilă au fost eficiente, indicând faptul că distorsiunile adevărului de bază nu au fost necesare pentru antrenare. Această constatare este semnificativă deoarece înseamnă că modelul poate fi antrenat pe o gamă largă de seturi de date, chiar și pe cele fără distorsiuni reale. Prin utilizarea de perechi de imagini distorsionate și nedistorsionate, modelul poate învăța să corecteze distorsiunile în mod eficient, ceea ce îl face versatil și aplicabil pe scară largă.

6.6 Concluzii

Această lucrare prezintă o metodă de învățare profundă pentru corectarea distorsiunilor complexe, utilă pentru autocalibrare fără ținte de calibrare specifice. Modelul nostru de distorsiune, bazat pe un model de thin plate spline, gestionează atât distorsiuni globale, cât și locale, ceea ce îl face potrivit pentru o gamă largă de aplicații, inclusiv conducerea autonomă și alte domenii în care camerele sunt utilizate în medii dinamice.

Am generat două seturi de date utilizând distribuții de distorsiuni din lumea reală. Rețeaua noastră neuronală a redus în mod eficient distorsiunile reziduale, sarcinile auxiliare îmbunătățind performanța. Experimentele au demonstrat că este posibilă antrenarea rețelei fără acces la distorsiuni reale, permițând extinderea la seturi de date cu înregistrări paralele de imagini distorsionate și nedistorsionate. Această capacitate face ca metoda noastră să fie extrem de versatilă și aplicabilă într-o gamă largă de scenarii în care metodele tradiționale de calibrare pot fi insuficiente.

În general, abordarea noastră reprezintă un progres semnificativ în domeniul calibrării camerelor și al corectării distorsiunilor. Prin valorificarea învățării profunde și a sarcinilor auxiliare, am dezvoltat o soluție robustă și scalabilă care poate gestiona distorsiuni complexe în contexte reale. Lucrările viitoare ar putea explora sarcini auxiliare suplimentare și ar putea rafina în continuare modelul pentru a îmbunătăți performanța și a extinde aplicabilitatea acestuia la noi domenii și provocări.

Capitolul 7

Concluzii

Atunci când camerele sunt montate în spatele unor obiecte transparente, din cauza refracției luminii, imaginile vor fi deformate, distorsionate. Această problemă afectează sistemele avansate de asistență a șoferului, unde o cameră utilizată pentru detectarea lumii din jurul mașinii este adesea montată în spatele parbrizului. Algoritmii de viziune computerizată geometrică necesită un model de cameră precis, care să poată proiecta coordonatele lumii 3D în pixeli 2D. În prezența unui parbriz, acest model de cameră trebuie să ia în considerare și distorsiunile. Acestea sunt de obicei mari și foarte neliniare, având atât componente globale, cât și locale. În lucrarea noastră am studiat problema calibrării camerei în prezența suprafețelor transparente și refractive.

Am propus o metodă inițială de calibrare, prin care construim un model precis al distorsiunilor cauzate de obiectele transparente din calea optică (cap. 4 și cap. 5). Am ales o abordare bazată pe fizică: în loc să abstractizăm componentele sistemului camerei, le modelăm explicit, inclusiv suprafețele refractive. Urmărim traiectoria razelor de lumină individuale de la centrul camerei la razele din lumea 3D, luând în considerare schimbările de direcție la limitele materialelor transparente. Ne bazăm această decizie pe faptul că, în cazurile noastre de utilizare, avem informații despre componente, pe care le putem încorpora în modelele noastre bazate pe fizică.

În primul rând, am modelat suprafața neuniformă a unui obiect transparent cu o formă globală. Acest lucru poate acoperi cazuri de utilizare în care elemente fizice precum parbrizele pot fi fabricate numai până la anumite toleranțe și pot avea neregularități. Am modelat suprafața folosind funcții de bază radiale, pe care le putem utiliza pentru a estima distorsiunile la nivel local. Parametrii modelului – ponderile (amplitudinile) – ai funcțiilor de bază radiale

au fost estimați prin tehnici de optimizare, pe baza imaginilor unor ținte de calibrare cu tablă de șah. Modelul nostru a îmbunătățit semnificativ erorile pentru colțurile tabloului de șah, atât pe seturi de date sintetice, cât și reale. Am furnizat, de asemenea, o analiză a distorsiunilor observate, unde am identificat că acestea au o relație liniară directă cu adâncimea inversă a pixelilor.

În această metodă am formulat modelul prin definirea funcției de retroproiecție. Deoarece această funcție implică trasarea razelor prin mai multe suprafețe, ea are două limitări principale. Din cauza complexității modelului, inversarea acestuia pentru a formula proiecția înainte se poate face numai prin metode iterative. De asemenea, complexitatea ridicată poate fi un dezavantaj pentru cazurile de utilizare integrate, în care modelul camerei trebuie evaluat frecvent, iar numărul mare de calcule poate crește timpul de execuție. Lucrările viitoare ar trebui să se concentreze pe găsirea unei soluții la această problemă, prin căutarea unor modele mai simple, care pot capta în continuare principalele proprietăți ale acestor distorsiuni, dar care, în același timp, pot fi integrate în sisteme reale. Analiza distorsiunilor poate oferi un bun punct de plecare în această direcție.

După ce am propus un model pentru suprafața locală neuniformă a obiectului transparent, ne-am îndreptat atenția către forma globală pentru cazurile de utilizare în care aceasta este necunoscută. Am propus să modelăm forma globală ca un elipsoid, care este suficient de general pentru a aproxima forma diferitelor obiecte din regiunea de interes a camerei. Acest model este direct compatibil cu modelul de suprafață din lucrările noastre anterioare. Am constatat că minimizarea acestei probleme este dificilă, deoarece este insuficient constrânsă: parametrii diferiți (forme globale) pot duce la distorsiuni foarte similare ale imaginii. A fost adăugat un termen de regularizare, care a reușit să orienteze minimizarea în direcția corectă.

Principalele limitări ale metodei propuse de noi sunt legate de metodologia noastră de testare, deoarece aceasta a fost realizată numai pe date sintetice. Extinderea acesteia la date reale ar trebui să fie principalul obiectiv al lucrărilor viitoare. După dovedirea acesteia pe date reale, cele două cadre propuse - un model global elipsoidal cu un model local bazat pe funcția de bază radială - ar putea fi o problemă de cercetare viitoare interesantă.

În cele din urmă, în cap. 6 am propus o soluție pentru autocalibrare bazată pe învățarea profundă. Am utilizat un set de date de măsurători ale distorsiunilor din parbrize reale și am construit un set de date sintetic și unul din lumea reală prin eșantionarea distorsiunilor în jurul celor măsurate. Arhitectura noastră bazată pe rețele neuronale convoluționale poate prezice

distorsiunile pe baza unei singure imagini sau pe baza unei secvențe de 3 imagini fără constrângeri privind mediul în care se află mașina. Arhitectura include, de asemenea, segmentarea semantică sau fluxul optic ca sarcini auxiliare, despre care demonstrăm că ne pot îmbunătăți semnificativ rezultatele. Utilizăm din nou un model de distorsiune, care include atât o componentă globală, cât și una locală: folosim un model de thin plate spline cu o componentă afină suplimentară pentru a modela deformările imaginii. Au fost testate două funcții de eroare, una bazată pe distorsiunile de referință și o a doua bazată pe reconstrucția imaginii. Cele mai bune rezultate au fost obținute folosind combinația celor două, dar antrenarea bazată doar pe reconstrucția imaginii a oferit rezultate competitive.

Principala limitare a acestei metode este setul nostru de date: distorsiunile, deși sunt bazate pe măsurători reale ale distorsiunilor din parbrize, au fost aplicate sintetic imaginilor. Viabilitatea utilizării doar a funcției de eroare bazată pe reconstrucția imaginilor deschide posibilitatea unor lucrări viitoare. În loc să se utilizeze distorsiuni cunoscute, ar putea fi creată o configurație de înregistrare paralelă, cu o cameră distorsionată de parbriz și una fără distorsiuni. Funcția de eroare bazată pe reconstrucție poate fi extinsă, presupunând o estimare corectă a distorsiunii, sintetizând imaginea nedistorsionată pe baza celei distorsionate. Dovedirea acestei metode fără date generate sintetic ar putea deschide posibilitatea unor aplicații reale.

Bibliografie

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. `http://ceres-solver.org`. Accessed: 2024-06-10.

[2] A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari. A theory of multi-layer flat refractive geometry. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3346–3353. IEEE, 2012. 10

[3] K. Anjyo, J. P. Lewis, and F. Pighin. Scattered data interpolation for computer graphics. In *ACM SIGGRAPH 2014 Courses*, pages 1–69. Association for Computing Machinery, 2014. 11

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] J. Beck and C. Stiller. Generalized b-spline camera model. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 2137–2142. IEEE, 2018. 10

[6] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY, USA, 2006. 16

[7] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 26

[8] C. Bräuer-Burchardt and K. Voss. Automatic lens distortion calibration using single views. In *Mustererkennung 2000*, pages 187–194. Springer Berlin Heidelberg, 2000.

[9] C. Brauer-Burchardt and K. Voss. A new algorithm to correct fish-eye-and strong wide-angle-lens-distortion from single images. In *Proceedings 2001 International Conference on Image Processing*, volume 1, pages 225–228. IEEE, 2001.

[10] D. C. Brown. Decentering distortion of lenses. *Photogrammetric Engineering and Remote Sensing*, 32(3):444–462, 1966.

[11] M. D. Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2000. 11

[12] M. Cassidy, J. Mélou, Y. Quéau, F. Lauze, and J.-D. Durou. Refractive multi-view stereo. In *2020 International Conference on 3D Vision (3DV)*, pages 384–393. IEEE, 2020.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 13

[14] T. Choi, S. Yoon, J. Kim, and S. Sull. Noniterative generalized camera model for near-central camera system. *Sensors*, 23(11):5294, 2023. 10

[15] D. Claus and A. W. Fitzgibbon. A rational function lens distortion model for general cameras. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 213–219. IEEE, 2005. 9

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. IEEE, 2016.

[17] S. Derrien and K. Konolige. Approximating a single viewpoint in panoramic imaging devices. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings*, volume 4, pages 3931–3938. IEEE, 2000.

[18] F. Devernay and O. Faugeras. Straight lines have to be straight. *Machine vision and applications*, 13(1):14–24, 2001. 9

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

[20] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In

Proceedings of the IEEE International Conference on Computer Vision, pages 2758–2766. IEEE, 2015. 13

[21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16. PMLR, 2017. 25

[22] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–125–I–132. IEEE, 2001. 9

[23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 25, 26

[24] D. B. Gennery. Generalized camera calibration including fish-eye lenses. *International Journal of Computer Vision*, 68:239–266, 2006. 9

[25] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279. IEEE, 2017.

[26] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838. IEEE, 2019.

[27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`. 6

[28] M. D. Grossberg and S. K. Nayar. A general imaging model and a method for finding its parameters. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 108–115. IEEE, 2001. 10

[29] M. D. Grossberg and S. K. Nayar. The raxel imaging model and ray-based calibration. *International Journal of Computer Vision*, 61(2):119–137, 2005.

[30] A. F. Habib, M. Morgan, and Y.-R. Lee. Bundle adjustment with self–calibration using straight lines. *The Photogrammetric Record*, 17(100):635–650, 2002.

[31] T. Hanning. *High precision camera calibration*. Springer, 2011. 6

[32] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034. IEEE, 2015. 27

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. IEEE, 2016. 12, 27

[34] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 12

[35] J. Heikkila. Geometric camera calibration using circular control points. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1066–1077, 2000.

[36] T. J. Herbert. Calibration of fisheye lenses by inversion of area projections. *Applied optics*, 25(12):1875–1876, 1986.

[37] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 13

[38] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 448–456. JMLR.org, 2015.

[39] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2017–2025. Curran Associates, Inc., 2015. 13, 27

[40] J. Kannala and S. S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006. 9

[41] J. Kannala, S. S. Brandt, and J. Heikkilä. Self-calibration of central cameras from point correspondences by minimizing angular error. In A. Ranchordas, H. J. Araújo, J. M. Pereira,

and J. Braz, editors, *Computer Vision and Computer Graphics. Theory and Applications*, pages 109–122. Springer Berlin Heidelberg, 2009.

[42] J. Kim, C. Kim, S. Yoon, T. Choi, and S. Sull. Rbf-based camera model based on a ray constraint to compensate for refraction error. *Sensors*, 23(20):8430, 2023. 10

[43] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12, 28

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[45] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[47] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11817–11825. IEEE, 2019.

[48] S.-B. Lőrincz, S. Pável, and L. Csató. Single view distortion correction using semantic guidance. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, July 2019. 7, 24

[49] L. Ma, Y. Chen, and K. L. Moore. Rational radial distortion models of camera lenses with analytical solution for distortion correction. *International Journal of Information Acquisition*, 1(02):135–147, 2004.

[50] H. Martins, J. R. Birk, and R. B. Kelley. Camera models based on data from two calibration planes. *Computer Graphics and Image Processing*, 17(2):173–180, 1981. 10

[51] C. Mei and P. Rives. Single view point omnidirectional camera calibration from planar grids. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3945–3950. IEEE, 2007.

[52] B. Micusik and T. Pajdla. Estimation of omnidirectional camera model from epipolar geometry. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[53] P. Miraldo and H. Araujo. Calibration of smooth camera models. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2091–2103, 2012.

[54] S. Morinaka, F. Sakaue, J. Sato, K. Ishimaru, and N. Kawasaki. 3d reconstruction under light ray distortion from parametric focal cameras. *Pattern Recognition Letters*, 124:91–99, 2019.

[55] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[56] E. Olson. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407. IEEE, May 2011.

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8024–8035, 2019.

[58] S. Pável. An ellipsoid object model of the refraction surface. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI)*, volume 2650, pages 272–279. CEUR Workshop Proceedings, January 2020. 6, 21

[59] S. Pável, C. Sándor, and L. Csató. Distortion estimation through explicit modeling of the refractive surface. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing*, pages 17–28. Springer, September 2019. 6, 15

[60] L. Qin, Y. Hu, Y. Wei, Y. Zhou, and H. Wang. Approach for camera self-calibration based on five straight lines. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1–4. IEEE, 2008.

[61] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436. IEEE, 2020.

[62] J. Rong, S. Huang, Z. Shang, and X. Ying. Radial lens distortion correction using convolutional neural networks trained with synthesized images. In S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, editors, *Computer Vision – ACCV 2016*, pages 35–49, Cham, 2017. Springer International Publishing.

[63] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 13

[64] D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, 1987.

[65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 27

[66] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520. IEEE, 2018.

[67] M. Schönbein, T. Strauß, and A. Geiger. Calibrating and centering quasi-central catadioptric cameras. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4443–4450. IEEE, 2014.

[68] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4168–4176. IEEE Computer Society, 2016.

[69] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *The 3rd International Conference on Learning Representations*, pages 1–15, 2015.

[70] P. Sturm and S. Ramalingam. A generic concept for camera calibration. In T. Pajdla and J. Matas, editors, *Computer Vision - ECCV 2004*, pages 1–13. Springer, Springer Berlin Heidelberg, 2004.

[71] P. Sturm, S. Ramalingam, J.-P. Tardif, S. Gasparini, J. Barreto, et al. Camera models and fundamental concepts used in geometric computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(1–2):1–183, 2011. 6

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. IEEE, 2015.

[73] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 6

[74] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 6105–6114. PMLR, 2019.

[75] T. Thormählen and H. Broszio. Automatic line-based estimation of radial lens distortion. *Integrated Computer-Aided Engineering*, 12(2):177–190, 2005.

[76] J. Tischendorf, C. Trautwein, T. Aach, D. Truhn, T. Stehle, et al. Camera calibration for fish-eye lenses in endoscopywith an application to 3d reconstruction. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1176–1179. IEEE, 2007.

[77] R. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, 1987.

[78] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Deep image harmonization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2799–2807, 2017.

[79] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[80] F. Verbiest, M. Proesmans, and L. Van Gool. Modeling the effects of windshield refraction for camera calibration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 397–412. Springer, 2020. 10

[81] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 13, 27

[82] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1369–1373. IEEE, 2015. 13

[83] Y. Wu and K. He. Group normalization. In *Computer Vision – ECCV 2018*, pages 3–19. Springer International Publishing, 2018.

[84] Y. Xiong and K. Turkowski. Creating image-based vr using a self-calibrating fisheye lens. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 237–243. IEEE, 1997.

[85] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao. Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In *Computer Vision – ECCV 2018*, pages 475–490. Springer International Publishing, 2018. 13

[86] S. Yoon, T. Choi, and S. Sull. Depth estimation from stereo cameras through a curved transparent medium. *Pattern Recognition Letters*, 129:101–107, 2020. 10

[87] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. `https://D2L.ai`. 6

[88] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22:1330–1334, 2000. 9, 18

[89] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[90] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858. IEEE, 2017. 13