

Doktori (PhD) értekezés tézisei

**Képi adatszintézisen alapuló mélytanuló algoritmusok
alkalmazása ipari robotikai rendszerek számára**

Kapusi Tibor Péter

Témavezető: Prof. Dr. habil Hajdu András



DEBRECENI EGYETEM

Informatikai Tudományok Doktori Iskola

Debrecen, 2026

Tartalomjegyzék

1. Téma terület bemutatása és a disszertáció célkitűzései	3
2. Az értekezés új tudományos eredményei és a tézisek	4
2.1. Képből-kép fordítás alkalmazása	4
2.2. Adatkészlet-fordítási módszer	6
2.3. A SCARA Assembly AI algoritmus	11
2.4. A többlépcsős adatszintetizáló modul	18
2.5. A PhD értekezés alapjául szolgáló közlemények	22
2.6. További közlemények	22
3 Introduction to the thesis topic and objectives	28
4 New scientific results and theses of the dissertation	29
4.1 Application of image-to-image translation	29
4.2 The dataset translation method	31
4.3 The SCARA Assembly AI algorithm	36
4.4 Multi-stage data synthesis module	43
4.5 List of publications related to the dissertation	47
4.6 List of other publications	47

1. Téma terület bemutatása és a disszertáció célkitűzései

A digitalizáció megjelenésével a megfelelő mennyiségű és információ-tartalmú adatgyűjtés megléte és biztosítása szintén meghatározó tényezővé válik, és ez az ipari modernizáció folyamatában (Ipar 4.0 és Ipar 5.0) is lényeges szerephez fog jutni. Ahogy az élet más területein is megfigyelhető, a gyártási folyamatokban is egyre hangsúlyosabban jelennek meg a digitális látáson és objektumazonosításon alapuló eljárások, miközben prioritássá válik a robotizált termelés, az automatikus anyagmozgatás és logisztika, valamint az intelligens szerelési rendszerek fejlesztése.

A disszertáció a fentebb említett gyártási folyamatok digitalizációjához és modernizációjához kapcsolódó kulcsfontosságú kutatási irányokat tárgyalja. A kutatás téma területa pedig az ipari mélytanuláson és szintetikus adatgeneráláson alapuló megoldások és eljárásokhoz kapcsolódik.

Kidolgozásra kerültek képtartomány transzformáción alapuló megközelítések és újfajta módszertanok is. Legelőször egy újfajta alkalmazását ismertetem a képből kép alapú fordítási eljárásnak ipari robotikai rendszerekben, amely lehetővé teszi komplexebb esetekben is a roboteszköz vizuális látásának biztosításához szükséges mélytanuló detektorok megfelelő pontosságú tanítását is.

Ezt követően folytatásként egy újfajta képi adatkészlet fordító eljárás megvalósítását írom le, amely az úgynevezett komplex wavelet alapú struktúralis hasonlóságon alapulva képes az adott gyártási környezet virtuális leképezésére, és valóság-hű felvételek szintetizálására ezzel jelentősen csökkentve a renderelt és valós képek közötti eltérést. Az eljárást valós eszközön és környezetben is teszteltem. Végül pedig egy olyan mélytanuláson alapuló eljárást készítettem, amely alkalmas ipari gyártási folyamatok megtanulására, és sorrendfüggetlen vezérlési szekvenciát biztosít SCARA típusú régi robotegység számára.

2. Az értekezés új tudományos eredményei és a tézisek

2.1. Képből-kép fordítás alkalmazása

Első tézispont

Tématerületnek megfelelően kidolgoztam egy saját optimalizált gépi tanulási architektúrát, amely képből-kép fordítási eljárásnak egy újfajta alkalmazását ipari robotika rendszerek számára. A megvalósításhoz a pix2pixHD alapú mélytanuló architektúrát vettem kiindulási alapként, és azon végeztem el architektúrális optimalizálást. A kiértékelés során megállapítottam, hogy az eljárás abban az esetben alkalmazható, ha a renderelt és valós képpárok rendelkezésre állnak. Megállapítottam továbbá azt is, hogy az elfogadható minőségű képszintetizáláshoz legalább 0,5-ös értékű komplex wavelet struktúrális hasonlóság szükséges, melynek értékét elvégzett kísérletek alapján határoztam meg.

Kapcsolódó publikáció:

[J3]. **Kapusi, T.P.; Erdei, T.I.; Husi, G.; Hajdu, A.** *Application of image translation-based approach in industrial manufacturing system for object identification. Int. Rev. Appl. Sci. Eng., 2025, 1–11. Q2*

Ebben a munkában többrétegű Gauss-képpiramis alapú konvolúciós zajszűrés és automatikus kódolási neurális hálózati modellt használtam. Ebben az esetben tartomány átalakítást kell alkalmazni, amely a $G : A \mapsto B$ képtartomány leképezés megtanulásán alapul (lásd a 1. ábrát).

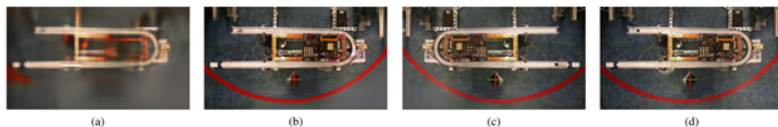
A legnagyobb kihívást a hálózat hiperparamétereinek legjobb kombinációjának megtalálása jelentette. Ezért folyamatosan tanítottam az architektúrát az adatkészletek alapján, és több architektúraparamétert is módosítottam. A diszkriminátorok számát 2-ről 5-re növeltük, a lefelé mintavételi szinteket



1. ábra. A nyers renderelt képek tartományfordítása.

3-ról 9-re emeltem, valamint a jellemzőklasztereket 10-ről 16-ra szintén módosítottam.

A 3.4 táblázat (lásd a disszertáció 3.5.2-es alfejezetében) a kiválasztott pix2pixHD módszer eredményeit tartalmazza. A 2. ábra a tanítási eljárás során kapott mintaképeket mutatja. További mérőszámokat is kiválasztottam, amelyeket az egész képre és az objektumok régióira alkalmaztam.



2. ábra. A pix2pixHD algoritmus kimenetei a szintetizált képsorozatok képzésének során: (a) mintavételezett kép a képzés kezdetén, (b) mintavételezett kép körülbelül 0,5-ös hasonlósági értéknél, (c) mintavételezett kép 0,6 feletti hasonlósági értéknél, (d) mintavételezett kép a képzés végén.

Az adatkészlet sztenderd értékeinek (legjobb és legrosszabb eredmények) kiszámításával szintén megállapítottam, hogy bár az objektumok szakaszai jelentősen nagyobb értékekkel rendelkeztek, a teljes képek nagyon kevés eltérést mutattak. Ebből kifolyólag a hasonlóság értéke nagy hatással lehet a detektorok teljesítményére a régiók minőségének arányától függően.

A 3.5 táblázat a betanított YOLO detektorok eredményeit tartalmazza (lásd a disszertáció 3.5.2-es alfejezetében). Referenciaként a detektorokat valós

képeken is betanítottam, és a táblázat első sora tartalmazza az ehhez tartozó értékeket. A generált adatkészlet viszonylag nagy volt, 256 000 feldúsított képből a tanításhoz és 400 valós képből állt a validáláshoz. Az előző fejezetben említettek szerint a kapott CW-SSIM érték döntő fontosságú a detektorok számára, amely technikai értelemben szintetikus küszöbértéknek is tekinthető. Az értékekből arra a következtetésre jutottunk, hogy amikor a komplex wavelet szerkezeti hasonlóság 0,5 alá esett, akkor a detektorok pontossága drámaian csökkent, mert az detektálandó objektum szinte teljesen beleolvadt a háttérbe.

2.2. Adatkészlet-fordítási módszer

Második tézispont

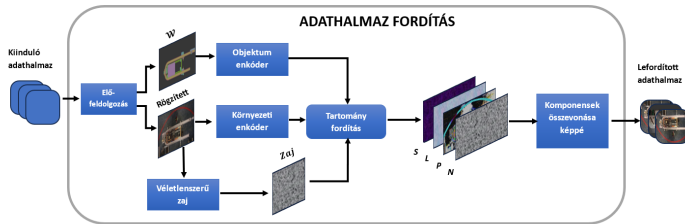
Kidolgoztam egy újfajta képi adatkészlet fordító eljárást, amely képes az adott gyártási környezetet leíró virtuális modell alapján valósághű képfelvételek szintetizálására. Figyelembe véve a szintetizált képi adathalmaz információmennyiségének hasonlóságát a valós felvételekhez képest megállapítottam, hogy legalább 0,6-os értékű egyezőségre van szükség a mélytanuló detektorok szinte teljesen azonos pontosságú működéséhez, az érték megállapítása elvégzett kísérletek alapján került meghatározásra.

Kapcsolódó publikáció:

[J2]. Erdei T.I.; **Kapusi, T.P.**; Hajdu, A.; Husi, G. *Image-to-Image Translation-Based Deep Learning Application for Object Identification in Industrial Robot Systems. Robotics*, 2024, **13**, 88. <https://doi.org/10.3390/robotics13060088> *Q1*

A javasolt adatkészlet-fordítási módszerünk felépítését és főbb összetevőit a 3. ábra mutatja be.

A képből-kép fordítás jelenleg számos mélytanulási architektúrával megva-



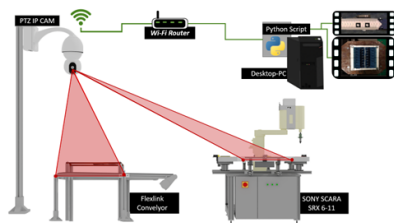
3. ábra. Az adatkészlet fordítási módszer architektúrája.

lósítható, amelyek közül a pix2pixHD-alapút választottuk kiinduló felépítésként, mivel kiváló pontossággal képes fotórealisztikus és nagy felbontású képeket generálni. A 4. ábra szemlélteti a generálási folyamatot, amely során a renderelt képeket konvertáljuk, és létrehozuk a detektorok betanításához szükséges képadatbázist.



4. ábra. Képjelenetek tartományfordítása.

A megvalósított tesztkörnyezetünket a 5. ábra szemlélteti. A különböző képzési adatkészletek alapján kapott eredményeket pedig a 1. táblázat tartalmazza. A tesztek elvégzéséhez a következő két detektort választottam ki számítási követelményeik és pontossági képességeik alapján: YOLOv3-tiny-3l a szállítószalag-adatkészlethez és YOLOv8-nano a Sony SCARA-adatkészlethez. A kapott eredmények alapján a szintetikusán generált adatkészleteken tanított detektorok közel olyan pontos detektálásokat képesek elérni, mintha csak valós képeken lettek volna tanítva (lásd a 6.



5. ábra. A megvalósított tesztkörnyezet.

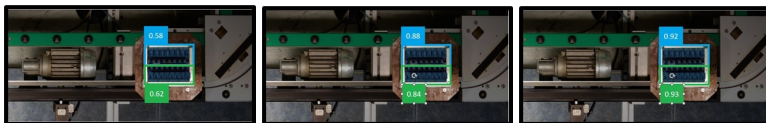
1. táblázat. A különböző tanítási adatkészletekkel kapcsolatos detektálási eredmények.

Training Dataset	Conveyor Dataset			Sony SCARA Dataset		
	Train	Val	Test	Train	Val	Test
Rendered only	0.624	0.612	0.511	0.689	0.674	0.655
Synthetic image translation generated	0.795	0.784	0.778	0.802	0.782	0.775
Real images	0.847	0.823	0.814	0.835	0.822	0.802

ábrát).

A tesztelés során mindkét eszközön kimértem a feldolgozás idejét is. A YOLOv3-tiny-3l esetében körülbelül 54,19 milliszekundumot, a YOLOv8-nano esetében pedig 76,29 milliszekundumot kaptunk, ami az Nvidia Jetson eszközön elfogadható eredménynek számít. Összehasonlításképpen a detektorokat egy asztali RTX 4060Ti grafikus kártyán lefuttatva átlagosan 27,87 és 73,11 milliszekundumot feldolgozási időket tapasztaltam.

A kiválasztott pix2pixHD algoritmus eredményeit a 4.5. és 4.6. táblázat tartalmazza (lásd a 4.9 alfejezet végén a disszertációban). A tanítási folyamat során kapott mintavételi képek a 7. ábrán láthatók. A táblázat további metrika értékeket is tartalmaz, a megfelelő szórás, valamint az egyes releváns tanítási ciklusok minimális és maximális értékeivel együtt. A kiértékelés alatt szintén a komplex hullám strukturális hasonlóság (CWSSIM)



Kizárólag renderelt képe- Szintetizált képeken beta- Kizárólag valós felvétele-
ken betanított. nított. ken betanított.

PCB-based Sony SCARA dataset



Kizárólag renderelt képe- Szintetizált képeken beta- Kizárólag valós felvétele-
ken betanított. nított. ken betanított.

V slot-based Conveyor dataset

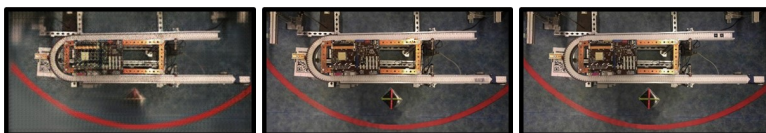
6. ábra. A pix2pixHD algoritmus tanítása során szintetizált képhalmazok kimenetei.

nevű robusztus algoritmust választottam, mert ez az algoritmus különböző transzformációk, például forgatás és átméretezés esetén is invariáns marad.



Eredmény a tanítás kezdé- Eredmény a tanítás köze- Eredmény a tanítási folya-
tén. pénél. mat végén.

PCB-based Sony SCARA dataset



Eredmény a tanítás kezdé- Eredmény a tanítás köze- Eredmény a tanítási folya-
tén. pénél. mat végén.

V slot-based Conveyor dataset

7. ábra. A pix2pixHD algoritmus tanítása során szintetizált képsorok kime-
netei.

2.3. A SCARA Assembly AI algoritmus

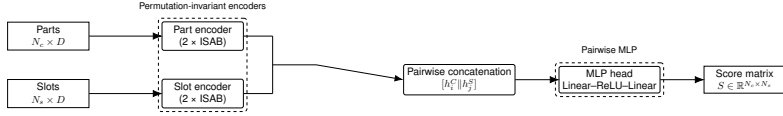
Harmadik tézispont

A SCARA Assembly AI nevű algoritmushoz kapcsolódóan kidolgoztam és implementáltam egy Set-transformers architektúrán alapuló mélytanuló neurális hálózatot, amely képes sorrend függetlenül az alkatrészek és a hozzátartozó célrégiók bijektív leképezésének megtanulására és megvalósítására. Kísérletileg igazoltam, hogy különböző méretű és komplexitású nyomtatott áramköri lapkák esetében az architektúra képes legalább 91%-os pontosság elérésére. A kidolgozott háromlépcsős mélytanuló architektúrán alapuló eljárás pedig képes sorrend függetlenül dinamikusan megtanulni és megvalósítani SCARA típusú roboteszközöket vezérelve ipari beültetési folyamatokat ellátni. Az algoritmus működésének pontosságát kísérletileg igazoltam, egy az Omniverse ISAAC Sim applikációban felépített virtuális környezetben. A kidolgozott módszertan validálása valós ipari környezetben is megtörtént SONY SCARA SRX-611 gépegységen, ahol a sikerességi szerelési arány 98% -os értéket ért el.

Tézishez kapcsolódó publikáció: [J1]. Kapusi, T.P.; Erdei, T.I.; Abdullah, M.; Husi, G.; Hajdu, A. SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly. Robotics, 2025, 14, 175. <https://doi.org/10.3390/robotics14120175> Q1

A slot-hozzárendelési feladat megoldásához Set-Transformer alapú architektúrát alkalmaztunk. A bemenetet a korábban említett két halmaz (komponensek és slotok) adja, amelyek elemei $p_i = [x, y, w, h, c] \in \mathbb{R}^5$ ötelemű jellemzővektorokkal írhatóak le (lásd az 8. ábrán).

Az összesített kimenet egy $S \in \mathbb{R}^{N_c \times N_s}$ dimenziós pontszámmátrix lesz; ez fogja kifejezni a slotok helykiosztásainak erősségét, amelyet a Hungarian



8. ábra. A hozzárendelő neurális hálózat felépítése [J1].

párosító algoritmus dolgoz fel a betanítási és az értékelési folyamat során.

A feladat elvégzéséhez optimalizáljuk a modell által létrehozott $S \in \mathbb{R}^{N_c \times N_s}$ pontszámmátrixot. Az $\mathcal{M}$ valós adatokat figyelembe véve a soft Hungarian loss-t a következőképpen határozzuk meg:

$$\mathcal{L}_{\text{Hungarian_basic}}(S, \mathcal{M}) = -\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \log \sigma(s_{ij}) \quad (1)$$

ahol $S = [s_{ij}] \in \mathbb{R}^{N_p \times N_s}$ egy pontszámmátrix, $\mathcal{M} \subseteq \{1, \dots, N_c\} \times \{1, \dots, N_s\}$ az alapadatok párpainak halmaza, $\sigma(\cdot)$ pedig a szigmoid függvényt jelöli.

Az (1) egyenletben szereplő alapvető veszteség csak a valós párok pontszámainak növelését ösztönzi, azonban nem bünteti explicit módon a hozzárendelés által kiválasztott, de helytelen párokat, illetve azokat a valós párokat, amelyeket a hozzárendelés kihagy. Ennek kezelésére először meghatározzuk a Hungarian algoritmus által adott hozzárendelést. Mivel a Hungarian algoritmus klasszikusan minimalizálási feladatot old meg, a maximalizálandó pontszámok helyett a negatívjukat adjuk meg költségként: $M_h = \text{Hungarian}(-S)$. Itt M_h a kiválasztott párok halmaza, $M_h \subseteq \{1, \dots, N_c\} \times \{1, \dots, N_s\}$, és egy-egy hozzárendelést reprezentál (egy index legfeljebb egy párban szerepelhet mindkét oldalon). Mivel a feladatban minden komponens legfeljebb egy foglalathoz, és minden foglalat legfeljebb egy komponenshez tartozik, a valós párok halmaza $\mathcal{M}_{gt}$ szintén 1:1 hozzárendelést reprezentál, ezért $|M_h| = |\mathcal{M}_{gt}| = K$, ahol

$$K = \min(N_c, N_s).$$

A kiválasztott és a valós párok viszonyát három, egymástól diszjunkt halmazzal írjuk le: $A = M_h \cap \mathcal{M}_{gt}$ (helyes választások), $B = M_h \setminus \mathcal{M}_{gt}$ (kiválasztott, de helytelen párok), és $C = \mathcal{M}_{gt} \setminus M_h$ (kihagyott valós párok). Fontos megjegyezni, hogy a C halmaz pontosan azokat a valós párokat tartalmazza, amelyek nem szerepelnek a Hungarian által kiválasztott M_h halmazban; ez nem „kétszeresen kihagyott” párokat jelent, hanem egyszerű halmazkülönbséget.

A kiválasztott párok minőségét bináris keresztentrópiával értékeljük úgy, hogy az A elemei pozitív, a B elemei negatív címkét kapnak:

$$\mathcal{L}_{\text{match}}(S, M_h, \mathcal{M}_{gt}) = -\frac{1}{|M_h|} \left(\sum_{(i,j) \in A} \log \sigma(s_{ij}) + \sum_{(i,j) \in B} \log(1 - \sigma(s_{ij})) \right). \quad (2)$$

A kihagyott valós párok külön büntetőtagot kapnak:

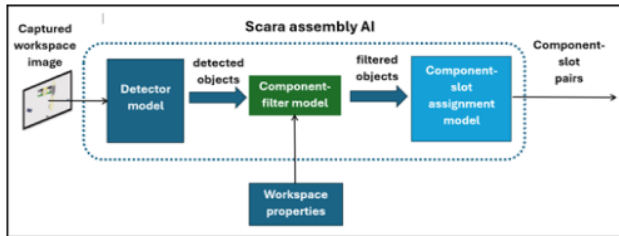
$$\mathcal{L}_{\text{miss}}(S, M_h, \mathcal{M}_{gt}) = -\frac{1}{|\mathcal{M}_{gt}|} \sum_{(i,j) \in C} \log \sigma(s_{ij}). \quad (3)$$

A végső veszteség a két tag súlyozott összege:

$$\mathcal{L}_{\text{Hungarian_assembly}} = \mathcal{L}_{\text{match}}(S, M_h, \mathcal{M}_{gt}) + \alpha \mathcal{L}_{\text{miss}}(S, M_h, \mathcal{M}_{gt}), \quad \alpha \geq 0. \quad (4)$$

Az általam javasolt és kidolgozott többszintű megközelítés felépítését szemlélteti az 9. ábra. A környezetben megtalálható objektumok felismerését egy valós idejű működésre képes objektumdetektor fogja ellátni. A komponens-foglalat hozzárendelést pedig egy Set-transformerekre alapuló egyedi neurális hálózat fogja biztosítani. Mivel az objektumdetektor esetében előfordulhatnak duplikált vagy téves detektálások, mindenképp szükség lesz egy szűrőkomponensre, amely képes lesz az objektumdetektor kimenetét szétválogatni, és csak a helyes adatokat meghagyni, ugyanis

mindenfajta hibás detektálás zajként fog megjelenni és megzavarhatja a hozzárendelési háló működését.



9. ábra. A javasolt módszerünk, az úgynevezett SCARA Assembly AI algoritmus architektúrája.

A kidolgozott módszertan egyik legfontosabb lépése a bemutatás és tesztelés, amely az Omniverse környezetben lett megvalósítva. Fő célom az volt, hogy a betanított architektúra kimenetét integráljam egy olyan környezetbe, ahol a robotkar szimulált módon valósághű mozdulatokat hajthat végre (lásd a 10. ábrát).



10. ábra. A járműgyártási laboratórium és kutatóközpont virtuális laboratóriumi helyszíne.

Következő lépésként a SCARA Assembly AI algoritmusának működése

került beépítésre, amelynek a főbb lépéseit az alábbi pszeudokód mutatja be:

Algorithm 1: SCARA Assembly AI algoritmus

Input: Jelenet JSON fájlok, konfiguráció szabályrendszerek, betanított modellek súlyfájljai
Mozgási szekvencia szögekben

Inicializáció:

$jelenet \leftarrow loadJSON()$
 $detektor, szuro, hozzarendelo \leftarrow modellekBetoltese(sulyok)$

Folyamat:

$objectumok \leftarrow detektor(jelenet);$ // Objektum
detektalas
 $helyes \leftarrow filter(objectumok);$ // Helytelen elemek
szures
 $parok \leftarrow hozzrendel(helyes, foglalatok);$
// Komponens-foglalat hozzárendelés

Inverz kinematika:

foreach (c, s) **in** $prok$ **do**
 $szogek \leftarrow IK(c.pozicio, s.pozicio)$
 $szekvencia.hozza fuzes(szogek)$

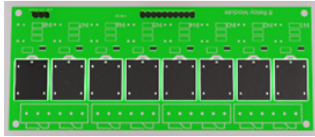
Kimenet:

A szekvencia visszaadása a környezetnek

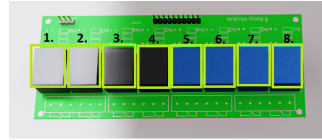
Az algoritmus által vizualizált hozzárendelési eredményeket a 11. ábracsoport szemlélteti.

Az összeállított valós tesztkörnyezet felépítését szemlélteti a 25 ábra. A munkaterület fölé 1,5 m magasságban egy Tapo C200 RGB kamerát rögzítettünk, amelyet merőlegesen igazítottunk a szerelési felülethez.

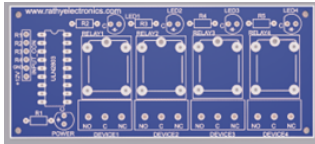
A 2. táblázat az összesített kiértékelést mutatja, és a teljes SCARA Assemb-



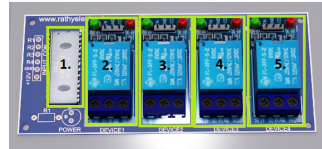
(a)



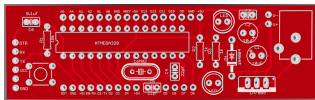
(b)



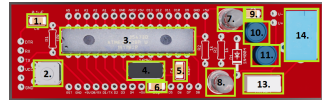
(c)



(d)

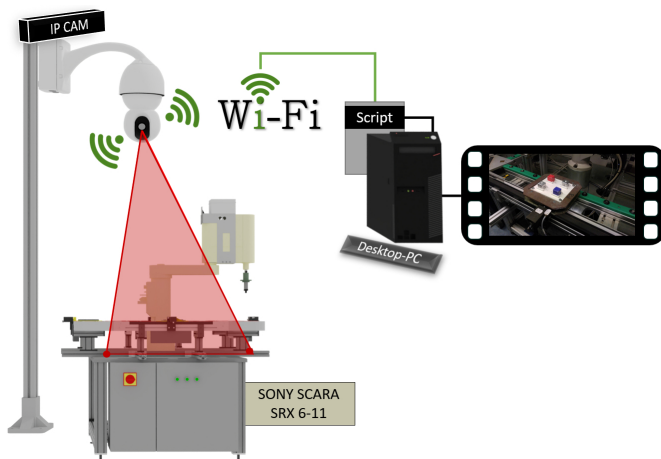


(e)



(f)

11. ábra. A SCARA Assembly AI algoritmus eredményei: (a,c,e) Az előre meghatározott foglalatokkal rendelkező üres PCB-k, (b,d,f) A PCB-k beültetett változata kijelölt alkatrészekkel.



12. ábra. A valós környezetben végzett validációhoz használt SCARA-alapú szerelőállomás kísérleti összeállítása. Az IP-kamera rögzíti a komponens-elhelyezési képeket, és Wi-Fi-n keresztül továbbítja azokat a vezérlő- és adatfeldolgozó szkripteket futtató asztali számítógépre. A SONY SCARA SRX-611 a detektált komponens–foglatat párok alapján hajtja végre a pick-and-place műveleteket.

ly AI algoritmus következetesen magas teljesítményt nyújt valamennyi észlelési és döntéshozatali modulban. A detektor szintetikus és valós körülmények között egyaránt 97,9%-os átlagos pontosságot ér el, míg a szűrőmodul a domain-randomizált vizuális zaj ellenére is 97,4%-os pontosságot tart fenn.

Modul	Környezet	Kulcsmetrika	Teljesítmény	Megjegyzés
Detektálás (YOLOv8s-Seg)	Omniverse és valós	Pontosság	97.9%	Stabil domain transzfer mellett
Szűrőhálózat	Csak szintetikus	Pontosság	97.4%	Robusztus vizuális zajjal szemben
Hozzárendelő hálózat	Szintetikus és valós	Pontosság	96.9%	Legjobb $\alpha = 1.0$ konfigurációban
Szimulációs inferencia	Omniverse	Késleltetés/átbocsátás	16 ms / ≈ 62 FPS	Valós idejű működésre képes
Valós környezeti teszt	SCARA SRX-611	Sikerráta	98%	Fizikai validáció robot hardveren

2. táblázat. A SCARA Assembly AI keretrendszer összesített kiértékelésének összefoglalója a detektálási, szűrési, hozzárendelési és valós környezeti validációs fázisokon keresztül.

2.4. A többlépcsős adatszintetizáló modul

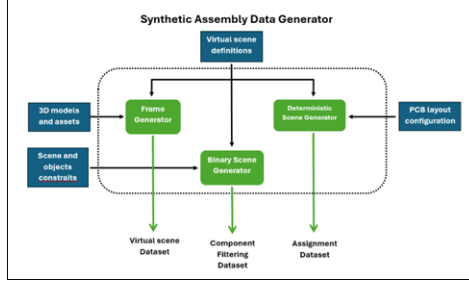
Negyedik tézispon

Megterveztem egy újfajta adatszintetizáló szoftvermodult, amely az NVIDIA által fejlesztett Omniverse platformon alapulva képes való-sághű adatkészletek előállítására és ipari robotika rendszerek környe-zetének és komponenseinek lemodellezésére. Kísérletileg igazoltam a szintetizált adatkészletek alkalmasak mélytanuló architektúrák betanítá-sára.

*Tézishez kapcsolódó publikáció: [J1]. **Kapusi, T.P.**; Erdei, T.I.; Masuk, A.; Husi, G.; Hajdu, A. SCARA Assembly AI: The synthetic learning-based method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly. Q1 Submitted*

Az 13. ábra a szintetikus adatgenerálás folyamatának fő lépéseit mutatja be, és több generátort is tartalmaz.

Jelölje $\mathcal{W} \subset \mathbb{R}^2$ a fizikai munkaterületet. Most definiáljuk annak virtuális megfelelőjét $\mathcal{W}_{virtual} = (\mathcal{R}, \mathcal{O})$ szintetikus jelenetként, ahol $\mathcal{R}$ jelöli az előre meghatározott régiók halmazát $\mathcal{W} \subset \mathbb{R}^2$ munkaterületen belül, és $\mathcal{O} = (o_1, \dots, o_2)$ lesz a modellekhez és transzformációkhoz társított, példányosított 3D objektumok halmaza.



13. ábra. A megvalósított szintetikus összeszerelési adatgenerátor (SADG) modul.

A $\text{Render} : \mathcal{W}_{\text{virtual}} \times \Theta \longrightarrow \{(I_t, A_t)\}_{t=1}^T$ renderelő függvény leképezi a szintetikus jelenetet és a θ kamera-megvilágítási paramétereket egy I_t halmaznyi RGB-képre, megfelelő A_t annotációkkal (b_i határolókeretekkel és l_i osztálycímkékkel). Az így kapott $\mathcal{D}_{\text{scenes}} = \{(I_t, A_t)\}_{t=1}^T$ adatkészletet a $\mathcal{W} \subset \mathbb{R}^2$ munkaterület virtuális analóg adatkészletének nevezzük. A 3. táblázat tartalmazza a detektor komponens eredményeit, amelyet a $\mathcal{D}_{\text{scenes}} = \{(I_t, A_t)\}_{t=1}^T$ adatkészlet alapján tanítottunk be.

3. táblázat. A detektor modell tanítási és validációs eredményei.

Method name	Train dataset	Val dataset	Test dataset
mAP50	0.9982	0.9982	0.9963
mAP50–95	0.9795	0.9640	0.9640
Precision	0.9865	0.9704	0.9655
Recall	0.9597	0.9467	0.9303

Legyenek a $\mathcal{C} = (c_1, c_2, \dots, c_K)$ komponenstípusok adottak, ahol mindegyik c_K a (w_k, h_k) fizikai méretekhez kapcsolódik.

Minden c_K komponenstípus esetében tartozik egy elfogadható elhelyezési

régiókból álló $\mathcal{R}_{c_k} = \{r_{k,1}, r_{k,2}, \dots, r_{k,m_k}\}$ halmaz előre definiálva, ahol $r_{k,j}$ egy tengelyre igazított határoló régiót jelöl a $\mathcal{W} \subset \mathbb{R}^2$ munkaterületen belül.

A BSG által generált x_i mintát az adott komponensstípus $x_i = [x, y, w, h, \text{label}_{\text{id}}, \text{IoU}_{\text{max}}] \in \mathbb{R}^6$ jellemzővektorának nevezzük, ahol (x, y) az objektum koordinátái, (w, h) az objektum dimenziói, label_{id} a komponens típusa, és IoU_{max} pedig a maximális átfedés lesz bármely más elhelyezett objektummal. A generátor kimenetét a $\mathcal{D}_{\text{bin}} = \{(x_i, y_i)\}_{i=1}^N$ komponenselemek bináris adatkészletének nevezzük, és a bináris szűrőhálózat betanításához alkalmazzuk.

A $\mathcal{D}_{\text{bin}} = \{(x_i, y_i)\}_{i=1}^N$ adatkészletre betanított szűrőháló eredményeit foglalja magába a 4. táblázat.

4. táblázat. A komponensszűrő modell tanítási és validálási eredményei.

Dataset name	Train loss	Val loss	Train Acc	Val Acc
Dataset A	0.0736	0.0905	0.966	0.955
Dataset B	0.0425	0.0720	0.979	0.957
Dataset C	0.1320	0.1470	0.923	0.893

Minden elemet egy $p_i = [x, y, w, h, c] \in \mathbb{R}^5$, $s_j = [x, y, w, h, c] \in \mathbb{R}^5$ ötdimenziós jellemzővektorral ábrázolunk, és a komponens-slot párosítás jellemzővektorainak nevezünk, ahol az (x, y) térbeli koordináták, (w, h) a dimenzióbeli méretek, és c a típusazonosító lesznek.

A fenti definíció esetében mivel a dummy objektumokat már kiszűrtük, így minden rész pontosan egy slot-hoz lesz hozzárendelve, ugyanis $\forall p_i \in \mathcal{P} \exists! s_j \in \mathcal{S} : (i, j) \in \mathcal{M}$.

Formálisan a $\mathcal{D}_{\text{asg}} = \{(X^{(b)}, Y^{(b)})\}_{b=1}^B$ adatkészletet a komponens-slot párosítás jellemzővektorait tartalmazó adatkészletének nevezzük, ahol min-

den minta $X^{(b)} = (\mathcal{P}^{(\ell)}, \mathcal{S}^{(\ell)})$, $Y^{(b)} = \mathcal{M}^{(\ell)}$ -ből áll.

A $\mathcal{D}_{\text{asg}} = \{(X^{(b)}, Y^{(b)})\}_{b=1}^B$ adatkészletre betanított hozzárendelési háló tanítási eredményeit tartalmazza a 5. táblázat három különböző esetre.

5. táblázat. A hozzárendelő modell tanítási és validációs eredményei.

Dataset name	Train loss	Val loss	Train Acc	Val Acc
Dataset A	0.1120	0.1430	1.000	1.000
Dataset B	0.1656	0.1824	0.9880	0.9780
Dataset C	0.1820	0.2435	0.9340	0.9134

Referált folyóiratszavak

2.5. A PhD értekezés alapjául szolgáló közlemények

- [J1] **Kapusi, T.P.**; Erdei, T.I.; Abdullah, M.; Husi, G.; Hajdu, A. *SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly*. *Robotics*, 2025, **14**, 175. <https://doi.org/10.3390/robotics14120175> **Q1**
- [J2] Erdei, T.I.; **Kapusi, T.P.**; Hajdu, A.; Husi, G. *Image-to-Image Translation-Based Deep Learning Application for Object Identification in Industrial Robot Systems*. *Robotics*, 2024, **13**, 88. <https://doi.org/10.3390/robotics13060088> **Q1**
- [J3] **Kapusi, T.P.**; Erdei, T.I.; Husi, G.; Hajdu, A. *Application of image translation-based approach in industrial manufacturing system for object identification*. *Int. Rev. Appl. Sci. Eng.*, 2025, 1–11. **Q2**

2.6. További közlemények

- [J4] Szántó, A., Abdullah, M., Kapusi, T., Diós, S. S. *Dynamic Modelling and Simulation of a Permanent Magnet Synchronous Motor (PMSM) Applied in a Prototype Race Car and the Comparison of Its Performance with BLDC Motor*. *Modelling*, 6 (3), 104, 1–21, 2025, **Q2**.
- [J5] Kapusi, T., Erdei, T. I., Husi, G., Hajdu, A. *Application of Deep Learning in the Deployment of an Industrial SCARA Machine for Real-Time Object Detection*. *Robotics*, 11 (4), 69, 1–20, 2022, **Q1**.
- [J6] Kapusi, T., Kovács, L., Hajdu, A. *Deep learning-based anomaly detection for imaging in autonomous vehicles*. In: *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, IEEE, Piscataway (NJ), 142–147, 2022. ISBN 978-1-6654-9652-0.
- [J7] Józsa, R. O., Erdei, T. I., Kapusi, T., Tóth, S., Husi, G. *MP-9S Industrial Manipulator Handling and Scrapping Tasks in Industrial*

- Environments. Műszaki Tudományos Közlemények*, 15 (1), 38–42, 2021.
- [J8] Tóth, J., Kapusi, T., Harangi, B., Tomán, H., Hajdu, A. *Accelerating the Optimization of a Segmentation Ensemble Using Image Pyramids*. In: *11th International Symposium on Image and Signal Processing and Analysis (ISPA 2019)*, IEEE, Piscataway (NJ), 43–48, 2019. ISBN 978-1-7281-3140-5.
- [J9] Krucsó, L., Erdei, T. I., Kapusi, T., Husi, G. *Designing an ATmega328 Microcontroller-Based Gesture-Controlled IoT UGV Unit and Creating a Camera System Using Linux Distribution*. *Recent Innovations in Mechatronics*, 6 (1), 1–7, 2019.
- [J10] Varga, T. A., Kapusi, T., Mankovits, T. *Celluláris anyagok modellezése digitális képek alapján / Modelling of Cellular Materials Based on Digital Images*. In: *XXII. Fiatal Műszakiak Tudományos Ülésszaka*, Erdélyi Múzeum Egyesület, Kolozsvár, 399–402, 2017. ISBN 978-963-449-018-0.
- [J11] Varga, T. A., Kapusi, T. *Digital Image Analysis of Metal Foam Specimens*. In: *4th International Scientific Conference on Advances in Mechanical Engineering (ISCAME 2016)*, University of Debrecen, Debrecen, 584–587, 2016. ISBN 978-963-473-944-9.



Nyilvántartási szám: DEENK/37/2026.PL
Tárgy: PhD Publikációs Lista

Jelölt: Kapusi Tibor Péter
Doktori Iskola: Informatikai Tudományok Doktori Iskola
MTMT azonosító: 10084802

A PhD értekezés alapjául szolgáló közlemények

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

1. **Kapusi, T.**, Erdei, T. I., Husi, G., Hajdu, A.: Application of image translation-based approach in industrial manufacturing system for object identification.
Int. rev. appl. sci. eng. [Epub ahead of print], 1-11, 2025. ISSN: 2062-0810.
DOI: <http://dx.doi.org/10.1556/1848.2025.01005>

Idegen nyelvű tudományos közlemények külföldi folyóiratban (2)

2. **Kapusi, T.**, Erdei, T. I., Abdullah, M., Husi, G., Hajdu, A.: SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly.
Robotics. 14 (12), 1-36, 2025. ISSN: 2218-6581.
DOI: <http://dx.doi.org/10.3390/robotics14120175>
IF: 3.3 (2024)
3. Erdei, T. I., **Kapusi, T.**, Hajdu, A., Husi, G.: Image-to-Image translation-based deep learning application to object identification in industrial robot systems.
Robotics. 13 (6), 1-21, 2024. ISSN: 2218-6581.
DOI: <http://dx.doi.org/10.3390/robotics13060088>
IF: 3.3

További közlemények

Idegen nyelvű, külföldi könyvrészletek (1)

4. **Kapusi, T.**, Kovács, L., Hajdu, A.: Deep learning-based anomaly detection for imaging in autonomous vehicles.
In: 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS): Proceedings . Szerk.: Fazekas István, IEEE, Piscataway (NJ), 142-147, 2022. ISBN: 9781665496520





Magyar nyelvű tudományos közlemények külföldi folyóiratban (1)

5. Józsa, R. O., Erdei, T. I., **Kapusi, T.**, Tóth, S., Husi, G.: MP-9S ipari manipulátor anyagmozgatási és selejtezési feladatok ellátása ipari környezetben.
Műsz. tud. közl. 15, 38-42, 2021. ISSN: 2393-1280.
DOI: <http://dx.doi.org/10.33895/mtk-2021.15.08>

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

6. Krucsó, L., Erdei, T. I., **Kapusi, T.**, Husi, G.: Designing an ATmega328 microcontroller based gesture-controlled IoT UGV unit and creating a camera system using Linux distribution.
Recent Innov. Mechatron. 6 (1), 1-7, 2019. EISSN: 2064-9622.

Idegen nyelvű tudományos közlemények külföldi folyóiratban (3)

7. Szántó, A., Abdullah, M., **Kapusi, T.**, Diós, S. S.: Dynamic Modelling and Simulation of a Permanent Magnet Synchronous Motor (PMSM) Applied in a Prototype Race Car and the Comparison of Its Performance with BLDC Motor.
Modelling. 6 (3), 1-21, 2025. EISSN: 2673-3951.
DOI: <http://dx.doi.org/10.3390/modelling6030104>
IF: 1.5 (2024)
8. **Kapusi, T.**, Erdei, T. I., Husi, G., Hajdu, A.: Application of Deep Learning in the Deployment of an Industrial SCARA Machine for Real-Time Object Detection.
Robotics. 11 (4), 1-20, 2022. ISSN: 2218-6581.
DOI: <https://doi.org/10.3390/robotics11040069>
IF: 3.7
9. Józsa, R. O., Erdei, T. I., **Kapusi, T.**, Tóth, S., Husi, G.: MP-9S Industrial Manipulator Handling and Scrapping Tasks in Industrial Environments. utánkölés nyelvi változat,
Műsz. tud. közl. 15 (1), 38-42, 2021. ISSN: 2601-5773.
DOI: <http://dx.doi.org/10.33894/mtk-2021.15.08>

Magyar nyelvű konferencia közlemények (1)

10. Varga, T. A., **Kapusi, T.**, Mankovits, T.: Celluláris anyagok modellezése digitális képek alapján / Modelling of Cellular Materials Based on Digital Images.
In: A XXII. Fiatal Műszakiak Tudományos Ülésszak előadásai = Proceedings of the XXIIth International Scientific Conference of Young Engineers. Szerk.: Bitay Enikő, Erdélyi Múzeum Egyesület (EME), Kolozsvár, 399-402, 2017. ISBN: 9789634490180

Idegen nyelvű konferencia közlemények (2)

11. Tóth, J., **Kapusi, T.**, Harangó, B., Tomán, H., Hajdu, A.: Accelerating the Optimization of a Segmentation Ensemble using Image Pyramids.
In: 11th International Symposium on Image and Signal Processing and Analysis (ISPA 2019). Eds.: S. Lončarić, R. Bregović, M. Carli, M. Subašić, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 43-48, 2019. ISBN: 9781728131405





12. Varga, T. A., **Kapusi, T.**: Digital image analysis of metal foam specimens.

In: Proceedings of the 4th International Scientific Conference on Advances in Mechanical Engineering (ISCAME 2016). Ed.: Sándor Bodzás, Tamás Mankovits, University of Debrecen Faculty of Engineering, Debrecen, 584-587, 2016. ISBN: 9789634739449

A közlő folyóiratok összesített impakt faktora: 11,8

**A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre):
6,6**

A DEENK a Jelölt által a Tudóstérbe feltöltött adatok bibliográfiai és tudánymetria ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2026.01.30.



Short thesis for the degree of doctor of philosophy (PhD)

**Application of deep learning algorithms based on image
data synthesis for industrial robotic systems**

by Tibor Péter Kapusi

Supervisor: Prof. Dr. habil András Hajdu



UNIVERSITY OF DEBRECEN

Doctoral School of Informatics

Debrecen, 2026

3 Introduction to the thesis topic and objectives

With the advent of digitalization, the existence and assurance of data collection in quantity and with sufficient information content is also becoming a decisive factor, and this will also play an important role in the process of industrial modernization (Industry 4.0 and Industry 5.0). As can be observed in other areas of life, processes based on digital vision and object identification are also appearing in manufacturing processes, and the development of robotized production, automatic material handling and logistics, and intelligent assembly systems is becoming a priority.

The dissertation discusses key research areas related to the digitization and modernization of the above-mentioned manufacturing processes.

The research topic is related to solutions and procedures based on industrial deep learning and synthetic data generation. Image domain transformation-based approaches and new methodologies will also be presented. First, I will describe a new application of image-to-image translation on industrial robotic systems, which allows for the accurate training of deep learning detectors necessary for ensuring the visual perception of robotic devices even in more complex cases. Next, I present a new type of image data set translation process based on complex wavelet-based structural similarity, which is capable of virtually mapping a given manufacturing environment and synthesizing realistic images. Finally, I describe a deep learning-based method that is capable of learning industrial manufacturing processes and providing sequence-independent control sequences for older SCARA-type robot units.

4 New scientific results and theses of the dissertation

4.1 Application of image-to-image translation

Thesis 1.

In line with the subject area, I developed my own optimized machine learning architecture, which is a new application of image-to-image translation for industrial robotics systems. For implementation, I took the pix2pixHD-based deep learning architecture as a starting point and performed architectural optimization on it. During the evaluation, I determined that the procedure can be applied if rendered and real image pairs are available. I also found that a complex wavelet structural similarity of at least 0.5 is required for acceptable image synthesis quality, which I determined based on experiments I conducted.

Related publication: [J3]. Kapusi, T.P.; Erdei, T.I.; Husi, G.; Hajdu, A. Application of image translation-based approach in industrial manufacturing system for object identification. Int. Rev. Appl. Sci. Eng., 2025, 1–11. Q2

In this work, I used a multi-layer Gaussian image pyramid-based convolutional noise filtering and automatic coding neural network model. This model is capable of synthesizing perception results by recreating image patches at a given resolution. In this case, domain transformation must be applied, which is based on learning the image domain mapping $G : A \mapsto B$ (see in Figure 14 for details).

The biggest challenge was finding the best combination of network hyper-parameters. Therefore, I continuously trained the architecture based on the datasets and modified several architecture parameters. We increased the number of discriminators from 2 to 5, raised the downsampling levels from



Figure 14: Domain translation of raw rendered images.

3 to 9, and also modified the feature clusters from 10 to 16.

Table 3.4 (see in Section 3.5.2) shows the results of the selected pix2pixHD method. Figure 15 shows the sample images obtained during the training process. I also selected additional metrics, which we applied to the entire image and to the regions of the objects.

By calculating the standard values (best and worst results) of the dataset, I also found that although the sections of the objects had significantly higher values, the complete images showed very little variation. In the second case, the similarity value can have a significant impact on detector performance, depending on the ratio of region quality.

Table 3.5 (see in Section 3.5.2) shows the results of the trained YOLO detectors. As a reference, I also trained the detectors on real images, and the first row of the table contains the corresponding values. The generated dataset was relatively large, consisting of 256,000 augmented images and 400 real images for validation. As mentioned in the previous chapter, the CW-SSIM value obtained is crucial for the detectors, which can be considered a synthetic threshold value in technical terms. From the values, I concluded that when the complex wavelet structural similarity fell below 0.5, the accuracy of the detectors decreased dramatically because the object to be detected merged into the background.

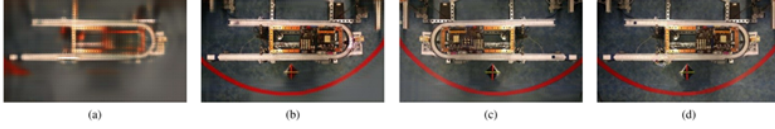


Figure 15: Outputs of the pix2pixHD algorithm during the training of synthesized image sequences: (a) sampled image at the beginning of training, (b) sampled image at a similarity value of approximately 0.5, (c) sampled image at a similarity value above 0.6, (d) sampled image at the end of training.

4.2 The dataset translation method

Thesis 2.

I have developed a new type of image data set translation method that is capable of synthesizing realistic images based on a virtual model describing a given manufacturing environment. Taking into account the similarity of the amount of information in the synthesized image data set compared to real images, I determined that a match of at least 0.6 is required for deep learning detectors to operate with almost identical accuracy. This value was determined based on the experiments performed.

Related publication: [J2].

Kapusi, T.P.; Hajdu, A.; Husi, G. *Image-to-Image Translation-Based Deep Learning Application for Object Identification in Industrial Robot Systems.* *Robotics*, 2024, **13**, 88. <https://doi.org/10.3390/robotics13060088> *Q1*

The structure and main components of our proposed dataset translation method are shown in Figure 16.

Image-to-image translation can currently be implemented using a number

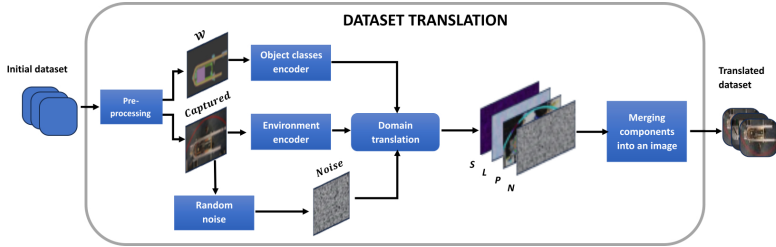


Figure 16: The architecture of dataset translation method.

of deep learning architectures, of which I chose pix2pixHD as our starting point because it is capable of generating photorealistic, high-resolution images with excellent accuracy. Figure 17 illustrates the generation process, during which I converted the rendered images and created the image database needed to train the detectors.



Figure 17: Image scene translation.

I also performed real-world tests with trained detectors on several devices, such as a desktop computer equipped with an Nvidia RTX 4060Ti GPU card and an Nvidia Jetson Nano development kit.

Table 6: Detection results for different training datasets.

Training Dataset	Conveyor Dataset			Sony SCARA Dataset		
	Train	Val	Test	Train	Val	Test
Rendered only	0.624	0.612	0.511	0.689	0.674	0.655
Synthetic image translation generated	0.795	0.784	0.778	0.802	0.782	0.775
Real images	0.847	0.823	0.814	0.835	0.822	0.802

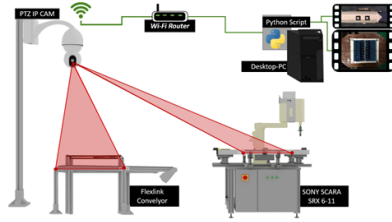
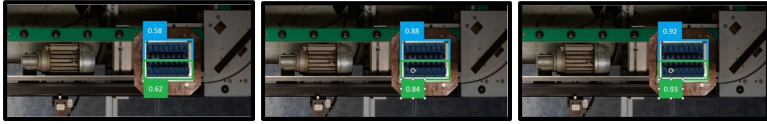


Figure 18: The implemented test environment.

Our test environment is illustrated in Figure 18. The results obtained from the different training datasets are shown in Table 19. I selected the following two detectors for the tests based on their computational requirements and accuracy: YOLOv3-tiny-3l for the conveyor belt dataset and YOLOv8-nano for the Sony SCARA dataset. Based on the results obtained, detectors trained on synthetically generated datasets can achieve nearly as accurate detections as those trained on real images (see in Figure 19).

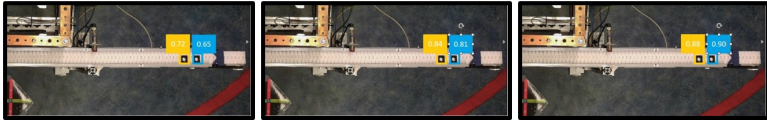
During testing, I also measured the processing time on both devices. I got approximately 54.19 milliseconds for YOLOv3-tiny-3l and 76.29 milliseconds for YOLOv8-nano, which is an acceptable result on the Nvidia Jetson device. Running the detectors on a desktop RTX 4060Ti graphics card, I achieved an average of 27.87 and 73.11 milliseconds.

The results of the selected pix2pixHD algorithm are shown in Tables 4.5 and 4.6 (see in Section 4.9). The sample images obtained during the training process are shown in Figure 20. The table also contains additional evaluation values, along with the corresponding standard deviation and the minimum and maximum values for each relevant epoch. During the evaluation, we also selected a robust algorithm called complex wavelet structural similarity (CWSSIM) because this algorithm is invariant in the case of various transformations, such as rotation and resizing, and is useful in our case.



Trained exclusively on rendered images. Trained on synthesized images. Trained exclusively on real images.

PCB-based Sony SCARA dataset



Trained exclusively on rendered images. Trained on synthesized images. Trained exclusively on real images.

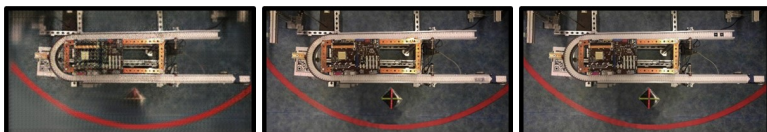
V slot-based Conveyor dataset

Figure 19: Outputs of image datasets synthesized during training of the pix2pixHD algorithm [J2].



Results at the beginning of the training. Results in the middle of the training. Result at the end of the training process.

PCB-based Sony SCARA dataset



Results at the beginning of the training. Results in the middle of the training. Result at the end of the training process.

V slot-based Conveyor dataset

Figure 20: Outputs of synthesized image sequences during training of the pix2pixHD algorithm.

4.3 The SCARA Assembly AI algorithm

Thesis 3.

In connection with the SCARA Assembly AI algorithm, I developed and implemented a deep learning neural network based on a Set-transformers architecture, which is capable of learning and implementing bijective mapping of parts and their associated target regions in a sequence-independent manner. I have experimentally verified that, for printed circuit boards of varying sizes and complexities, the architecture is capable of achieving at least 91% accuracy. The developed three-step deep learning architecture-based process is capable of dynamically learning and implementing SCARA-type robotic devices in any order to perform industrial implantation processes. I experimentally verified the accuracy of the algorithm in a virtual environment built in the Omniverse ISAAC Sim application. The developed methodology was also validated in a real industrial environment on a SONY SCARA SRX-611 machine unit, where the successful assembly rate reached 98%.

Relevant publication: [J1]. Kapusi, T.P.; Erdei, T.I.; Abdullah, M.; Husi, G.; Hajdu, A. SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly. Robotics, 2025, 14, 175. <https://doi.org/10.3390/robotics14120175> Q1

I used a Set-Transformer-based architecture to solve the slot assignment problem. The input is given by the two sets mentioned earlier (components and slots), whose elements can be described by five-element feature vectors $p_i = [x, y, w, h, c] \in \mathbb{R}^5$ (see Figure 21).

The aggregate output will be a $S \in \mathbb{R}^{N_c \times N_s}$ dimensional score matrix expressing the strength of the partial assignments processed by the Hungarian matching algorithm during the training and evaluation process.

To perform this task, I optimize the score matrix $S \in \mathbb{R}^{N_c \times N_s}$ created by the model. Taking into account the real data $\mathcal{M}$, I define the soft Hungarian loss as follows:

$$\mathcal{L}_{\text{Hungarian_basic}}(S, \mathcal{M}) = -\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \log \sigma(s_{ij}) \quad (5)$$

where $S = [s_{ij}] \in \mathbb{R}^{N_p \times N_s}$ is a score matrix, $\mathcal{M} \subseteq \{1, \dots, N_c\} \times \{1, \dots, N_s\}$ is the set of pairs of base data, and $\sigma(\cdot)$ denotes the sigmoid function.

However, the basic Hungarian loss in equation (5) does not penalize incorrect pairs selected by the assignment, nor does it penalize real pairs (GT) omitted by the assignment. To solve this problem, I extended the basic loss function and calculate it as follows: $M_h = \text{Hungarian}(-S)$.



Figure 21: The structure of the associative neural network.

Let $\mathcal{M}_{gt}$ be the set of GT pairs, and define three independent sets: $A = M_h \cap \mathcal{M}_{gt}$ (correct choices), $B = M_h \setminus \mathcal{M}_{gt}$ (selected but incorrect), and $C = \mathcal{M}_{gt} \setminus M_h$ (omitted GT pairs). I then use a binary cross-entropy-based matching term that rewards A and penalizes B , as well as an additional omission term that also penalizes C . The final goal will be to calculate their weighted sum using the following equations (6),(7),(8).

The results of the trained assignment module for datasets of varying complexity are shown in Table 10.

$$\mathcal{L}_{\text{match}}(S, M_h, \mathcal{M}_{gt}) = -\frac{1}{|M_h|} \left(\sum_{(i,j) \in A} \log \sigma(s_{ij}) + \sum_{(i,j) \in B} \log(1 - \sigma(s_{ij})) \right), \quad (6)$$

$$\mathcal{L}_{\text{miss}}(S, M_h, \mathcal{M}_{\text{gt}}) = -\frac{1}{|\mathcal{M}_{\text{gt}}|} \sum_{(i,j) \in \mathcal{C}} \log \sigma(s_{ij}), \quad (7)$$

$$\mathcal{L}_{\text{Hungarian_assembly}} = \mathcal{L}_{\text{match}}(S, M_h, \mathcal{M}_{\text{gt}}) + \alpha \mathcal{L}_{\text{miss}}(S, M_h, \mathcal{M}_{\text{gt}}), \quad \alpha \geq 0. \quad (8)$$

Figure 22 shows the structure of the multi-level approach I proposed and developed. The recognition of objects found in the environment will be performed by a real-time object detector. Component-slot assignment will be provided by a unique neural network based on Set Transformers. Since duplicate or false detections may occur in the case of the object detector, a filter component will be needed to sort the output of the object detector and retain only the correct data. This component is necessary because any kind of false detection will appear as noise and can interfere with the operation of the assignment network.

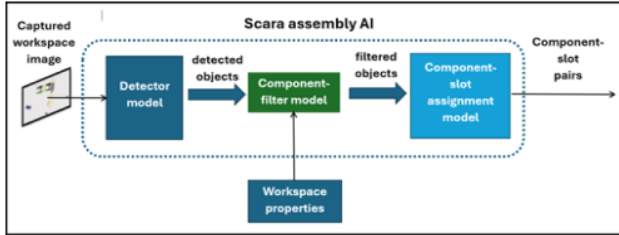


Figure 22: Our proposed method, the so-called SCARA Assembly AI algorithm architecture.

One of the most important steps in the developed methodology is presentation and testing. Accordingly, this section presents this part step by step. The test itself was simulated in the Omniverse environment. Our main goal was to integrate the output of the trained architecture into an environment

where the robotic arm can perform realistic movements in a simulated manner.



Figure 23: Virtual laboratory site of the vehicle manufacturing laboratory and research center.

The next step was to implement the SCARA Assembly AI algorithm (see the next pseudocode for details):

Algorithm 2: SCARA Assembly AI algorithm

Input: Scene JSON files, configuration rules, pretrained model weights

Output: Motion sequence

Initialization:

$scene \leftarrow loadJSON()$

$detector, filter, assigner \leftarrow loadModels(weights)$

Pipeline:

$objects \leftarrow detector(scene);$ // Object detection

$valid \leftarrow filter(objects);$ // Dummy filtering

$pairs \leftarrow assigner(valid, slots);$ // Component-slot assignment

Inverse kinematics:

foreach (c, s) **in** $pairs$ **do**

$angles \leftarrow IK(c.position, s.position)$

$sequence.append(angles)$

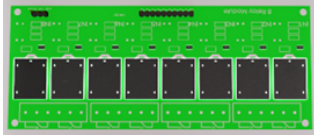
Output:

Return the sequence to the environment

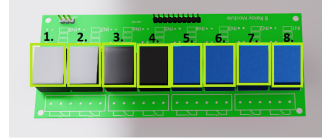
The assignment results visualized by the algorithm are illustrated in the group of Figures 24.

The structure of the compiled real test environment is shown in 25 figure. Above the work area 1,5 m We mounted a Tapo C200 RGB camera at a height of, which we aligned perpendicularly to the mounting surface.

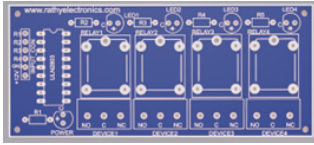
The table 7 shows the overall evaluation, and the complete SCARA Assembly AI algorithm consistently performs well in all perception and decision-making modules. The detector achieves an average accuracy of 97.9% in both synthetic and real-world conditions, while the Visual Filter Network maintains an accuracy of 97.4% despite domain-randomized visual noise.



(a)



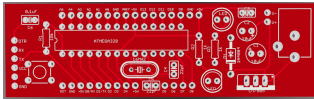
(b)



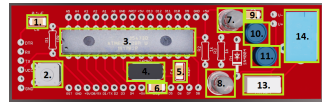
(c)



(d)



(e)



(f)

Figure 24: Results of the SCARA Assembly AI algorithm: (a,c,e) Empty PCBs with predefined slots, (b,d,f) PCBs with components installed in designated locations.

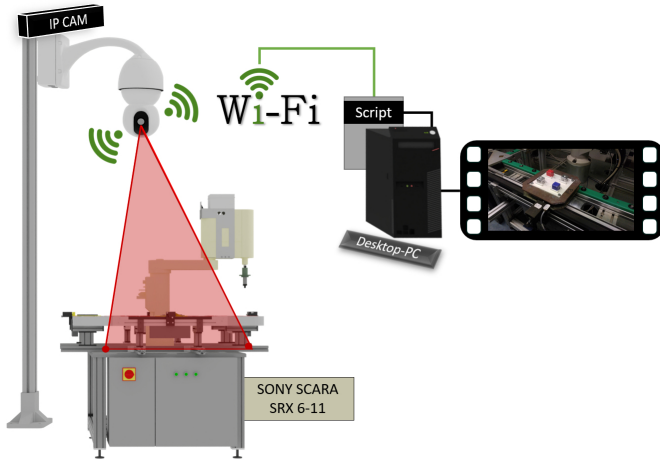


Figure 25: Experimental assembly of a SCARA-based assembly station used for validation in a real environment. The IP camera records images of component placement and transmits them via Wi-Fi to a desktop computer running control and data processing scripts. The SONY SCARA SRX-611 performs pick-and-place operations based on detected component-slot pairs.

Table 7: Overall evaluation summary of the SCARA Assembly AI framework across detection, filtering, assignment, and real-world validation stages. The integrated system achieves high accuracy ($\approx 97\%$) and real-time operation (≈ 16 ms per cycle), maintaining consistent performance across both simulated and physical environments.

Module	Environment	Key Metric	Performance	Remarks
Detection (YOLOv8s-Seg)	Omniverse and Real	Accuracy	97.9%	Stable under domain transfer
Filter Network	Synthetic only	Accuracy	97.4%	Robust to visual noise
Assignment Network	Synthetic and Real	Accuracy	96.9%	Best at $\alpha = 1.0$ configuration
Simulation inference	Omniverse	Latency/Throughput	16 ms/62 FPS	Real-time capable
Real-world test	SCARA SRX-611	Success rate	98%	Physical validation on robot hardware

4.4 Multi-stage data synthesis module

Thesis 4.

I designed a new type of data synthesis software module based on the Omniverse platform developed by NVIDIA, which is capable of generating realistic datasets and modeling the environment and components of industrial robotics systems. I experimentally verified that the synthesized datasets are suitable for training deep learning architectures.

Relevant publication: [J1]. **Kapusi, T.P.; Erdei, T.I.; Abdullah, M.; Husi, G.; Hajdu, A. SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly. *Robotics*, 2025, **14**, 175. <https://doi.org/10.3390/robotics14120175> Q1**

Figure 26 shows the main steps of the synthetic data generation process and includes several generators.

Let $\mathcal{W} \subset R^2$ denote the physical workspace. Now we define its virtual counterpart $\mathcal{W}_{virtual} = (\mathcal{R}, \mathcal{O})$ as a synthetic scene, where $\mathcal{R}$ denotes a set of predefined regions $\mathcal{W} \subset R^2$ within the workspace, and $\mathcal{O} = (o_1, \dots, o_n)$ is the set of instantiated 3D objects associated with the models and transformations.

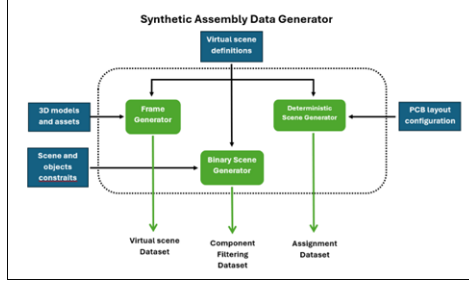


Figure 26: The implemented synthetic assembly data generator (SADG) module.

The Render : $\mathcal{W}_{\text{virtual}} \times \Theta \longrightarrow \{(I_t, A_t)\}_{t=1}^T$ renders the synthetic scene and the θ camera-illumination parameters into a set of I_t RGB images with corresponding A_t annotations (b_i bounding boxes and l_i class labels). The resulting dataset $\mathcal{D}_{\text{scenes}} = \{(I_t, A_t)\}_{t=1}^T$ is called the virtual analog dataset of the workspace $\mathcal{W} \subset R^2$. Table 8 shows the results of the detector component, which was trained on the dataset $\mathcal{D}_{\text{scenes}} = \{(I_t, A_t)\}_{t=1}^T$.

Table 8: Detector model training and validation results.

Method name	Train dataset	Val dataset	Test dataset
mAP50	0.9982	0.9982	0.9963
mAP50-95	0.9795	0.9640	0.9640
Precision	0.9865	0.9704	0.9655
Recall	0.9597	0.9467	0.9303

Let $\mathcal{C} = (c_1, c_2, \dots, c_K)$ be given component types, where each c_K is associated with physical dimensions (w_k, h_k) .

For each component type c_K , there is a predefined set of acceptable placement regions $\mathcal{R}_{c_k} = \{r_{k,1}, r_{k,2}, \dots, r_{k,m_k}\}$ of acceptable placement re-

gions, where $r_{k,j}$ denotes an axis-aligned bounding region within the workspace $\mathcal{W} \subset \mathbb{R}^2$.

The x_i sample generated by the BSG is called the characteristic vector of the given component type $x_i = [x, y, w, h, \text{label}_{\text{id}}, \text{IoU}_{\text{max}}] \in \mathbb{R}^6$ of the given component type, where (x, y) are the coordinates of the object, (w, h) are the dimensions of the object, label_{id} is the component type, and IoU_{max} is the maximum overlap with any other placed object.

The output of the generator is called the binary data set of the component elements $\mathcal{D}_{\text{bin}} = \{(x_i, y_i)\}_{i=1}^N$, and is used to train the binary filter network.

Table 9 shows the results of the filter network trained on the dataset $\mathcal{D}_{\text{bin}} = \{(x_i, y_i)\}_{i=1}^N$.

Table 9: Training and validation results of the component filter model.

Dataset name	Train loss	Val loss	Train Acc	Val Acc
Dataset A	0.0736	0.0905	0.966	0.955
Dataset B	0.0425	0.0720	0.979	0.957
Dataset C	0.1320	0.1470	0.923	0.893

Each element is represented by a five-dimensional characteristic vector $p_i = [x, y, w, h, c] \in \mathbb{R}^5$, $s_j = [x, y, w, h, c] \in \mathbb{R}^5$ five-dimensional feature vector, and we call them component-slot pairing feature vectors, where (x, y) are spatial coordinates, (w, h) are dimensional sizes, and c is the type identifier.

In the above definition, I have already filtered out the dummy objects, so each part will be assigned to exactly one slot, since $\forall p_i \in \mathcal{P} \exists! s_j \in \mathcal{S} : (i, j) \in \mathcal{M}$.

Formally, the dataset $\mathcal{D}_{\text{asg}} = \{(X^{(b)}, Y^{(b)})\}_{b=1}^B$ is called the data set of

feature vectors of component-slot pairings, where each sample $X^{(b)} = (\mathcal{P}^{(\ell)}, \mathcal{S}^{(\ell)})$, $Y^{(b)} = \mathcal{M}^{(\ell)}$.

The training results of assignment network trained on the dataset $\mathcal{D}_{\text{asg}} = \{(X^{(b)}, Y^{(b)})\}_{b=1}^B$ summarized in Table 10 for three different cases.

Table 10: Training and validation results of the assignment model.

Dataset name	Train loss	Val loss	Train Acc	Val Acc
Dataset A	0.1120	0.1430	1.000	1.000
Dataset B	0.1656	0.1824	0.9880	0.9780
Dataset C	0.1820	0.2435	0.9340	0.9134

Referred own publications

4.5 List of publications related to the dissertation

- [J1] **Kapusi, T.P.**; Erdei, T.I.; Abdullah, M.; Husi, G.; Hajdu, A. *SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly*. *Robotics*, 2025, **14**, 175. <https://doi.org/10.3390/robotics14120175> **Q1**
- [J2] Erdei, T.I.; **Kapusi, T.P.**; Hajdu, A.; Husi, G. *Image-to-Image Translation-Based Deep Learning Application for Object Identification in Industrial Robot Systems*. *Robotics*, 2024, **13**, 88. <https://doi.org/10.3390/robotics13060088> **Q1**
- [J3] **Kapusi, T.P.**; Erdei, T.I.; Husi, G.; Hajdu, A. *Application of image translation-based approach in industrial manufacturing system for object identification*. *Int. Rev. Appl. Sci. Eng.*, 2025, 1–11. **Q2**

4.6 List of other publications

- [J4] Szántó, A., Abdullah, M., Kapusi, T., Diós, S. S. *Dynamic Modelling and Simulation of a Permanent Magnet Synchronous Motor (PMSM) Applied in a Prototype Race Car and the Comparison of Its Performance with BLDC Motor*. *Modelling*, 6 (3), 104, 1–21, 2025 **Q2**.
- [J5] Kapusi, T., Erdei, T. I., Husi, G., Hajdu, A. *Application of Deep Learning in the Deployment of an Industrial SCARA Machine for Real-Time Object Detection*. *Robotics*, 11 (4), 69, 1–20, 2022 **Q1**.
- [J6] Kapusi, T., Kovács, L., Hajdu, A. *Deep learning-based anomaly detection for imaging in autonomous vehicles*. In: *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, IEEE, Piscataway (NJ), 142–147, 2022. ISBN 978-1-6654-9652-0.
- [J7] Józsa, R. O., Erdei, T. I., Kapusi, T., Tóth, S., Husi, G. *MP-9S Industrial Manipulator Handling and Scrapping Tasks in Industrial*

- Environments. Műszaki Tudományos Közlemények*, 15 (1), 38–42, 2021.
- [J8] Tóth, J., Kapusi, T., Harangi, B., Tomán, H., Hajdu, A. *Accelerating the Optimization of a Segmentation Ensemble Using Image Pyramids*. In: *11th International Symposium on Image and Signal Processing and Analysis (ISPA 2019)*, IEEE, Piscataway (NJ), 43–48, 2019. ISBN 978-1-7281-3140-5.
- [J9] Krucsó, L., Erdei, T. I., Kapusi, T., Husi, G. *Designing an ATmega328 Microcontroller-Based Gesture-Controlled IoT UGV Unit and Creating a Camera System Using Linux Distribution*. *Recent Innovations in Mechatronics*, 6 (1), 1–7, 2019.
- [J10] Varga, T. A., Kapusi, T., Mankovits, T. *Celluláris anyagok modellezése digitális képek alapján / Modelling of Cellular Materials Based on Digital Images*. In: *XXII. Fiatal Műszakiak Tudományos Ülésszaka*, Erdélyi Múzeum Egyesület, Kolozsvár, 399–402, 2017. ISBN 978-963-449-018-0.
- [J11] Varga, T. A., Kapusi, T. *Digital Image Analysis of Metal Foam Specimens*. In: *4th International Scientific Conference on Advances in Mechanical Engineering (ISCAME 2016)*, University of Debrecen, Debrecen, 584–587, 2016. ISBN 978-963-473-944-9.



Registry number:
Subject:

DEENK/37/2026.PL
PhD Publication List

Candidate: Tibor Péter Kapusi

Doctoral School: Doctoral School of Informatics

MTMT ID: 10084802

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (1)

1. **Kapusi, T.**, Erdei, T. I., Husi, G., Hajdu, A.: Application of image translation-based approach in industrial manufacturing system for object identification.

Int. rev. appl. sci. eng. [Epub ahead of print], 1-11, 2025. ISSN: 2062-0810.

DOI: <http://dx.doi.org/10.1556/1848.2025.01005>

Foreign language scientific articles in international journals (2)

2. **Kapusi, T.**, Erdei, T. I., Abdullah, M., Husi, G., Hajdu, A.: SCARA Assembly AI: The Synthetic Learning-Based Method of Component-to-Slot Assignment with Permutation-Invariant Transformers for SCARA Robot Assembly.

Robotics. 14 (12), 1-36, 2025. ISSN: 2218-6581.

DOI: <http://dx.doi.org/10.3390/robotics14120175>

IF: 3.3 (2024)

3. Erdei, T. I., **Kapusi, T.**, Hajdu, A., Husi, G.: Image-to-Image translation-based deep learning application to object identification in industrial robot systems.

Robotics. 13 (6), 1-21, 2024. ISSN: 2218-6581.

DOI: <http://dx.doi.org/10.3390/robotics13060088>

IF: 3.3

List of other publications

Foreign language international book chapters (1)

4. **Kapusi, T.**, Kovács, L., Hajdu, A.: Deep learning-based anomaly detection for imaging in autonomous vehicles.

In: 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS):

Proceedings . Szerk.: Fazekas István, IEEE, Piscataway (NJ), 142-147, 2022. ISBN:

9781665496520





Hungarian scientific articles in international journals (1)

5. Józsa, R. O., Erdei, T. I., **Kapusi, T.**, Tóth, S., Husi, G.: MP-9S ipari manipulátor anyagmozgatási és selejtezési feladatok ellátása ipari környezetben.
Műsz. tud. közl. 15, 38-42, 2021. ISSN: 2393-1280.
DOI: <http://dx.doi.org/10.33895/mtk-2021.15.08>

Foreign language scientific articles in Hungarian journals (1)

6. Krucsó, L., Erdei, T. I., **Kapusi, T.**, Husi, G.: Designing an ATmega328 microcontroller based gesture-controlled IoT UGV unit and creating a camera system using Linux distribution.
Recent Innov. Mechatron. 6 (1), 1-7, 2019. EISSN: 2064-9622.

Foreign language scientific articles in international journals (3)

7. Szántó, A., Abdullah, M., **Kapusi, T.**, Diós, S. S.: Dynamic Modelling and Simulation of a Permanent Magnet Synchronous Motor (PMSM) Applied in a Prototype Race Car and the Comparison of Its Performance with BLDC Motor.
Modelling. 6 (3), 1-21, 2025. EISSN: 2673-3951.
DOI: <http://dx.doi.org/10.3390/modelling6030104>
IF: 1.5 (2024)
8. **Kapusi, T.**, Erdei, T. I., Husi, G., Hajdu, A.: Application of Deep Learning in the Deployment of an Industrial SCARA Machine for Real-Time Object Detection.
Robotics. 11 (4), 1-20, 2022. ISSN: 2218-6581.
DOI: <https://doi.org/10.3390/robotics11040069>
IF: 3.7
9. Józsa, R. O., Erdei, T. I., **Kapusi, T.**, Tóth, S., Husi, G.: MP-9S Industrial Manipulator Handling and Scrapping Tasks in Industrial Environments. utánkölés nyelvi változat,
Műsz. tud. közl. 15 (1), 38-42, 2021. ISSN: 2601-5773.
DOI: <http://dx.doi.org/10.33894/mtk-2021.15.08>

Hungarian conference proceedings (1)

10. Varga, T. A., **Kapusi, T.**, Mankovits, T.: Celluláris anyagok modellezése digitális képek alapján / Modelling of Cellular Materials Based on Digital Images.
In: A XXII. Fiatal Műszakiak Tudományos Ülésszak előadásai = Proceedings of the XXII-th International Scientific Conference of Young Engineers. Szerk.: Bitay Enikő, Erdélyi Múzeum Egyesület (EME), Kolozsvár, 399-402, 2017. ISBN: 9789634490180

Foreign language conference proceedings (2)

11. Tóth, J., **Kapusi, T.**, Harangi, B., Tomán, H., Hajdu, A.: Accelerating the Optimization of a Segmentation Ensemble using Image Pyramids.
In: 11th International Symposium on Image and Signal Processing and Analysis (ISPA 2019). Eds.: S. Lončarić, R. Bregović, M. Carli, M. Subašić, Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 43-48, 2019. ISBN: 9781728131405





12. Varga, T. A., **Kapusi, T.**: Digital image analysis of metal foam specimens.

In: Proceedings of the 4th International Scientific Conference on Advances in Mechanical Engineering (ISCAME 2016). Ed.: Sándor Bodzás, Tamás Mankovits, University of Debrecen Faculty of Engineering, Debrecen, 584-587, 2016. ISBN: 9789634739449

Total IF of journals (all publications): 11,8

Total IF of journals (publications related to the dissertation): 6,6

The Candidate's publication data submitted to the Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

30 January, 2026

