Volume XI Issue 3 Year 2026 | Page 706-716 | ISSN: 2527-9866
Received: 11-05-2026 | Revised: 15-06-2026 | Accepted: 01-08-2026

Implementation of U-Net for Semantic Segmentation and Perspective Transformation in Floor Tile Virtual Try-On Systems

Fahrul Rozi[1], Ayu Ratna Juwita[2], Yusuf Eka Wicaksana[3], Deden Wahiddin[4], Sutan Faisal[5], Dwi Sulistya Kusumaningrum[6]
[1,2,3,4,5,6] Universitas Buana Perjuangan Karawang, Karawang, West Java, Indonesia, 41361
e-mail: if22.fahrulrozi@mhs.ubpkarawang.ac.id[1], ayurj@ubpkarawang.ac.id[2], yusuf.eka@ubpkarawang.ac.id[3], deden.wahiddin@ubpkarawang.ac.id[4], sutan.faisal@ubpkarawang.ac.id[5], dwi.sulistya@ubpkarawang.ac.id[6]
**Correspondence:* if22.fahrulrozi@mhs.ubpkarawang.ac.id

Abstract: *Conventional floor tile VTO systems rely on printed catalogs, physical samples, or marker-based AR, making it difficult for consumers to visualize ceramic products in real rooms. This study proposes a markerless VTO framework integrating U-Net floor segmentation and homography-based perspective transformation. The model was pre-trained on 5,285 SUN RGB-D images and fine-tuned on 1,338 customer-room images from PT Concord Industry, divided into training, validation, and independent test sets using a 70:20:10 ratio. Segmentation was evaluated using IoU, Dice Coefficient, Pixel Accuracy, Precision, and Recall. Perspective alignment was evaluated using four-corner Mean Angular Error (MAE) on two room images representing conditions with and without excessive furniture, while system functions were validated through black-box testing. The fine-tuned U-Net with a ConvNeXt-V2-Huge encoder achieved an IoU of 0.9483, a Dice coefficient of 0.9671, a pixel accuracy of 0.9891, a precision of 0.9675, and a recall of 0.9702. The perspective module achieved an average MAE of 5.92° and produced visually coherent tile alignment. These results demonstrate the framework's potential to support realistic floor tile visualization and reduce uncertainty during product selection.*

Keywords: *floor tile, markerless system, perspective transformation, semantic segmentation, u-net, virtual try-on.*

1. Introduction

Floor tiles have developed from surface coverings into important interior elements whose motifs, colors, and patterns influence the aesthetics, ambiance, and visual comfort of a property [1]. This role is increasingly relevant to the Indonesian ceramic industry. The Ministry of Industry through BSKJI reports that the national ceramic industry has an installed production capacity of 625 million m^2, while its utilization increased to 75% in the first quarter of 2025 from around 60% in 2024 [2]. These official figures are used to describe the industrial context that motivates digital product visualization.

Computer Vision (CV) supports automatic scene understanding and perspective manipulation for product visualization [3]. However, ceramic selection still commonly relies on printed catalogs and physical samples that provide limited spatial context. Empirical evidence on ceramic visual preferences indicates that visual product characteristics affect consumers' implicit evaluations [4]. This finding supports the need for tools that help consumers compare motifs in their own room context instead of mentally translating catalog images into a real interior scene.

Previous floor tile visualization research has mainly used marker-based AR [5], which achieved functional software performance but remained dependent on physical markers. Deep learning studies have improved indoor semantic segmentation [6][7][8], and homography estimation has been widely used for geometric view transformation [9]. Nevertheless, the integration of deep

learning-based floor segmentation with automatic geometric perspective transformation in a unified markerless floor tile VTO workflow remains insufficiently explored. This gap is important because floor replacement requires both accurate floor masks and viewpoint alignment; segmentation alone cannot make a tile motif follow the geometry of the room.

This study contributes: (1) a markerless VTO workflow that removes dependence on physical markers, (2) a two-stage transfer-learning strategy for automatic floor segmentation using U-Net, (3) automated homography-based perspective transformation for mapping ceramic motifs according to floor geometry and camera viewpoint, and (4) a web-based prototype that connects the segmentation and transformation modules into a usable VTO pipeline. Therefore, the research gap addressed in this study is the absence of an integrated markerless workflow for floor tile visualization that combines semantic segmentation, perspective correction, and web-based product interaction. The novelty lies in connecting these components into a single automated VTO system and evaluating both segmentation quality and perspective alignment.

2. Literature Review

Semantic segmentation is a deep learning technique that classifies every pixel in an image into specific class labels, providing a comprehensive understanding of image content at the pixel level [10]. Unlike object detection that produces only bounding boxes, semantic segmentation can precisely delineate object boundaries from the background, essential for handling objects with irregular shapes.

U-Net is a deep learning architecture specifically designed for precise image segmentation, using an encoder-decoder concept with a symmetric U-shaped structure that enables detailed spatial information recovery through skip connections [11]. Its main advantage lies in achieving high-accuracy segmentation even with relatively limited training datasets, making it highly effective for semantic segmentation tasks requiring very detailed object boundaries [12].

Perspective transformation is a geometric operation that projects an image from one viewpoint to a new viewpoint, mathematically represented using a 3x3 homography matrix [9]. This technique accounts for vanishing points to correct visual distortion when 3D objects are photographed into 2D images. In digital floor replacement applications, this transformation is required to manipulate flat ceramic motifs to match the tilt angle of the floor in the original photo, creating a realistic depth illusion [13]. VTO technology has rapidly expanded from the fashion industry to interior design sectors, where scale and perspective accuracy are determining factors for simulation success [14]. Research shows that VTO usage can increase user engagement and reduce customer hesitation by providing realistic and interactive product visualization [15].

Table 1. Summary of Related Research

No.	Researcher & Year	Method	Result
1	Basri et al., (2024) [5]	Marker-based AR Android App	ISO 25010: 100% functional, marker dependent
2	Ngoc et al., (2024) [7]	U-Net + scSE Attention	+12% mIoU vs baseline U-Net
3	Phu et al., (2025) [8]	U-Net + deep encoder backbone	More stable multi-class segmentation
4	Turkmen & Akgun (2025) [16]	Hybrid Ensemble U-Net	ARE=0.063, RMSE=0.237
5	Luo et al., (2023) [9]	Homography matrix estimation survey	Comprehensive review of methods

3.Methods

This study used an experimental research approach with six main stages: data collection, data preprocessing, segmentation model training, perspective transformation development, system implementation, and evaluation. The overall research flow is illustrated in Figure 1.

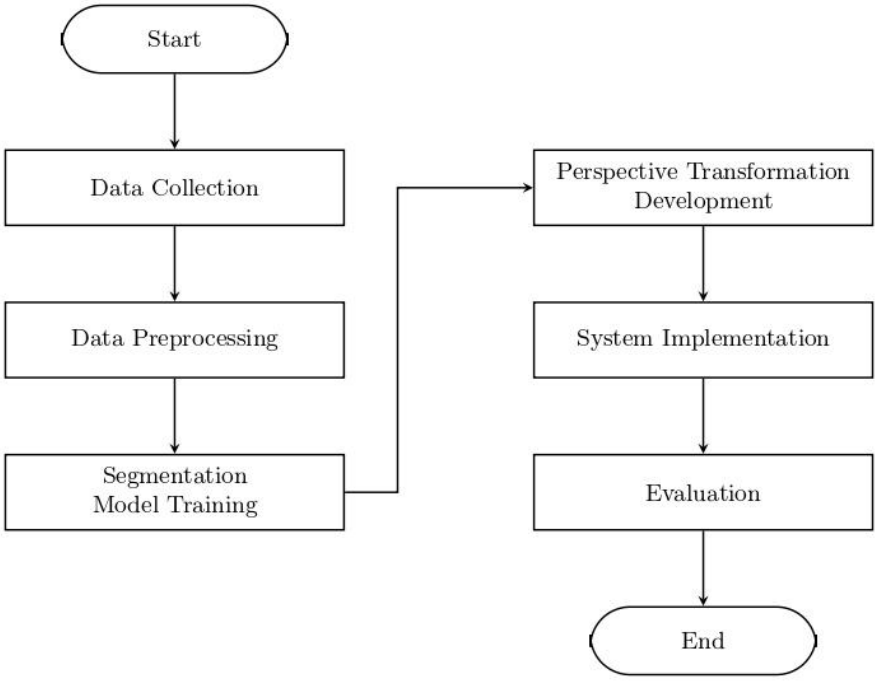

Figure 1. Research methodology flowchart

a. Data Collection.

Interior images were collected from PT Concord Industry customer rooms using a smartphone camera at a minimum resolution of 1920x1080 pixels, a camera height of 120-150 cm, a horizontal or slightly downward angle (0-30 degrees), and at least 30% visible floor area. Only room-interior and floor visual information was used for annotation and model training. The product assets comprised 50 Granite Tile motifs from the company's digital catalog at 512x512 pixels. Room images were manually annotated in Roboflow using floor polygons and exported as binary PNG masks with values 0 (non-floor) and 255 (floor). Annotation quality was checked through visual inspection of polygon continuity, floor boundaries, and mask-image correspondence before export; evident boundary errors were corrected. SUN RGB-D data obtained from the wyrx/SUNRGBD_seg repository were used for pre-training.

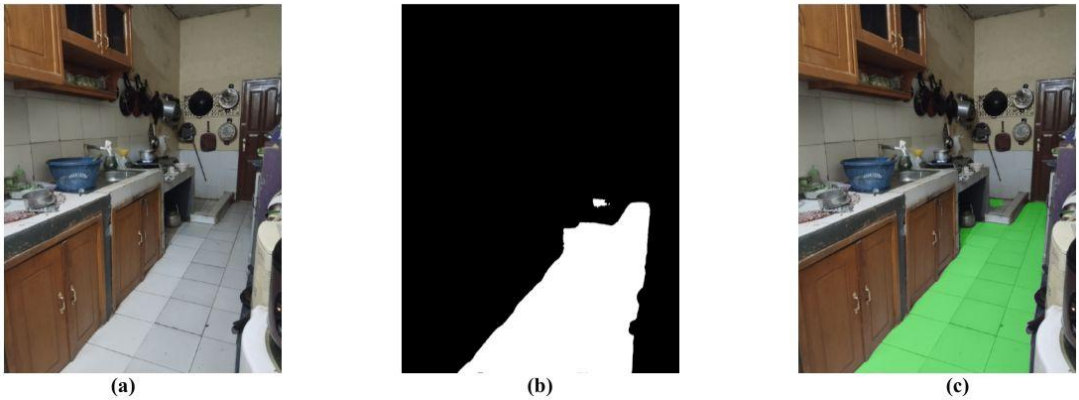

Figure 2. Local dataset sample: (a) Original room image, (b) Binary annotation mask, (c) Segmentation overlay

b. Data Preprocessing.

The preprocessing stage included: (1) Image normalization to the [0,1] range by dividing pixel values by 255; (2) Dimension standardization to 512x512 pixels using bilinear interpolation for RGB images and nearest-neighbor interpolation for segmentation masks; (3) On-the-fly data augmentation during training using Albumentations library [17], including horizontal flip (p=0.5), affine transform (p=0.4), random resized crop (p=0.4), grid distortion (p=0.2), color jitter (p=0.5), random brightness contrast (p=0.4), and hue saturation value adjustment (p=0.3); (4) Dataset splitting using stratified sampling at 80:20 ratio for pre-training and 70:20:10 for fine-tuning via Roboflow.

c. Segmentation Model Training.

U-Net used the ImageNet-pre-trained ConvNeXt-V2-Huge encoder (convnextv2_huge. fcmae_ft_in22k_in1k_512) in features-only mode, producing 352, 704, 1,408, and 2,816-channel feature maps. This large backbone was selected during initial development to prioritize robust feature extraction for complex indoor boundaries, thin floor edges, and partially occluded floor regions; its computational cost is acknowledged as a limitation for resource-constrained deployment. Three decoder blocks used 2x2 transposed convolution, attention-gated skip connections, two 3x3 convolutions, Batch Normalization, GELU, and Dropout2D (0.1).

Training was conducted in two stages. Pre-training used the SUN RGB-D dataset with batch size 16, AdamW optimizer (learning rate 1x10^-4, weight decay 0.05), and Cosine Annealing scheduler with Linear Warmup. Fine-tuning used the local dataset with a lower learning rate (1x10^-5) for 10 epochs. A Compound Loss combining Binary Cross Entropy (BCE), Dice Loss, and Focal Loss was used throughout training.

d. Perspective Transformation

The perspective transformation module processed the binary segmentation mask from the U-Net model through the following steps: (1) Contour detection using cv2.findContours to identify the outer boundary of the floor area; (2) Convex hull algorithm application on the selected contour; (3) Polygon approximation using cv2.approxPolyDP to simplify the polygon into four main corner points representing the floor area perspective; (4) Homography matrix H (3x3) computation using cv2.findHomography based on four correspondence point pairs; (5) Ceramic motif grid tiling followed by perspective transformation using cv2.warpPerspective.

e. System Implementation

The VTO system was implemented as a web-based application allowing users to upload room photos, select ceramic tile motifs from the PT Concord Industry digital catalog (50 Granite Tile motifs), and view VTO visualization results. The system was built using Python with Flask framework for the backend, with the U-Net model served as an inference engine, and a responsive HTML/CSS/JavaScript frontend for the user interface.

f. Evaluation Method. Evaluation Method.

Segmentation was measured using IoU, Dice Coefficient, Pixel Accuracy, Precision, and Recall on the designated independent test set. Stage-wise comparison between the pre-training and fine-tuning stages was used as an ablation-style analysis to observe the contribution of local-domain adaptation. Black-box testing covered all web-system functions. Perspective alignment was evaluated on frontal, left-skewed, and right-skewed images. Manual and predicted TL, TR, BL, and BR corner points were used to calculate the left and right floor-boundary convergence angles as follows:

$$MAE_i = \frac{1}{4}\left(\left|\theta_{TL,i}^{pred} - \theta_{TL,i}^{manual}\right| + \left|\theta_{TR,i}^{pred} - \theta_{TR,i}^{manual}\right| + \left|\theta_{BL,i}^{pred} - \theta_{BL,i}^{manual}\right| + \left|\theta_{BR,i}^{pred} - \theta_{BR,i}^{manual}\right| \right) \quad (1)$$

$$MAE_{avg} = \frac{1}{m}\sum_{i=1}^{m} MAE_i \quad (2)$$

The four corner points were used to form the homography, while MAE summarized the angular deviation of the left and right floor boundary lines rather than the pixel error of each individual corner point, where *m* denotes the number of viewpoint scenarios.

4. Results and Discussion

a. Data Collection Results.

The primary dataset collection resulted in 1,338 interior room images with a floor area ratio distribution of 30-70% of the total frame. All 1,338 images were manually annotated, producing binary segmentation masks. The secondary SUN RGB-D dataset provided 5,285 labeled images covering various room types including bedrooms, living rooms, kitchens, offices, and classrooms. The SUN RGB-D dataset was divided using a stratified split with a proportion of 80% for training (4,228 images) and 20% for validation (1,057 images). Stratification was performed based on the presence of floor areas in each image using the train_test_split function from scikit-learn with a random_state=42 parameter to ensure reproducibility. Stratification verification results showed that the distribution of images with and without floor areas was consistent between the training and validation sets.

For the local dataset (primary data), the division was carried out through the Roboflow platform, which automatically divided the data into training, validation, and testing sets before being exported and integrated for the fine-tuning process. Based on this allocation, 1,338 primary images were augmented to 4,014 images, distributed with a proportion of 70% as training data (2,811 images), 20% validation data (804 images), and 10% testing data (399 images). This proportional approach was applied to ensure the model received adequate training data volume for feature adaptation while maintaining an independent, representative dataset for validation and final performance evaluation.

b. Model Training Results.

Pre-training on the SUN RGB-D dataset ran for 34 epochs before being halted by the early stopping mechanism. The training curve showed a consistent training loss decrease from 0.5916 (epoch 1) to 0.0431 (epoch 34). Validation IoU progressively increased to a best value of 0.8102 at epoch 19.

Table 2. Pre-training Model Performance at Key Epoch Milestones

z	Train. Loss	Val. Loss	Val. IoU	Dice Coefficient
1	0.5916	0.4268	0.5997	0.7434
5	0.2473	0.2127	0.7973	0.8854
12	0.1273	0.1728	0.7994	0.8869
19	0.0808	0.1869	0.8102	0.8935
34	0.0431	0.2232	0.8087	0.8926

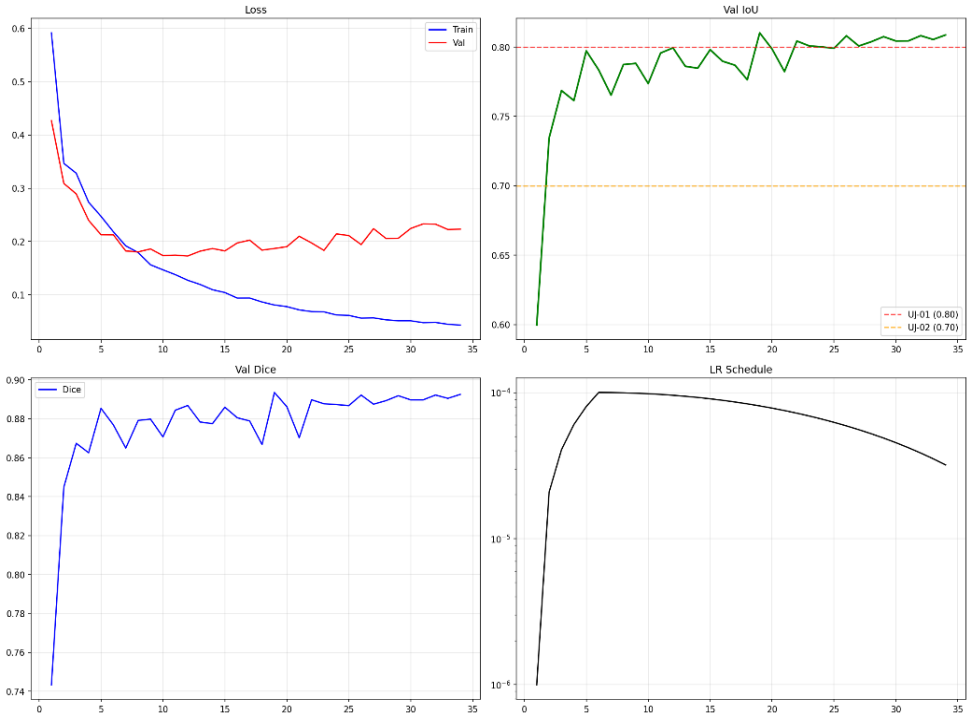

Figure 3. Pre-training learning curves Loss, Validation IoU, Dice, and LR Schedule

Fine-tuning reduced training loss from 0.0574 to 0.0264 and increased validation IoU from 0.9370 to 0.9483 at epoch 8. The comparison describes performance across the two transfer-learning stages; it is not a comparison against a separately trained encoder architecture. The +0.1381 IoU increase shows that local fine-tuning substantially adapted the model from general indoor scenes to the specific floor tile VTO domain.

Table 3. Comparison of Pre-training and Fine-tuning Metrics

Metric	Pre-training	Fine-tuning	Improvement
Validation IoU	0.8102	0.9483	+17.0%
Training Loss	0.0431	0.0264	-0.0167
Validation Loss	0.1728	0.0326	-0.1402
Total Epochs	34	10	-

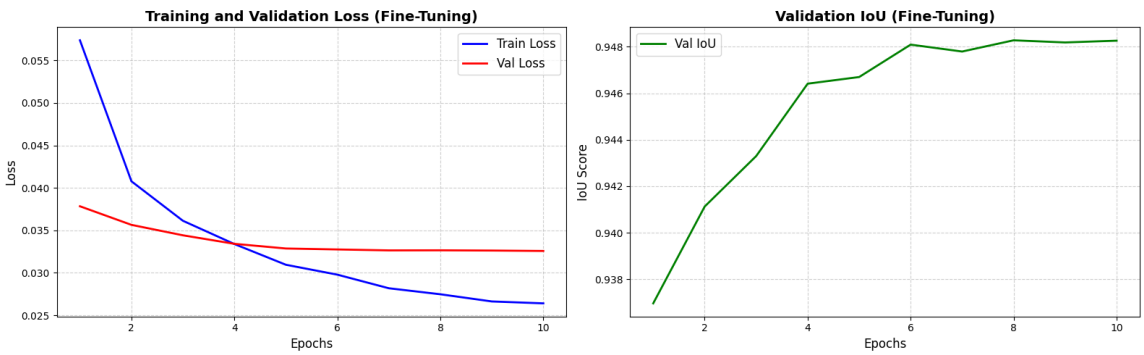

Figure 4. Fine-tuning learning curves Training/Validation Loss and Validation IoU

c. Model Evaluation Results.

To ensure objective performance measurement, the final model (post fine-tuning) was evaluated using standard semantic segmentation metrics strictly on the unseen independent testing dataset (399 images).

Table 4. Quantitative Evaluation Results of the U-Net Model

Metric	Value
Intersection over Union (IoU)	0.9483
Dice Coefficient	0.9671
Pixel Accuracy	0.9891
Precision	0.9675
Recall	0.9702

The test-set IoU of 0.9483 indicates substantial overlap between predictions and ground truth. The model also produced consistent floor masks across lighting variation, partial furniture occlusion, and complex floor geometry. Error analysis showed that false-positive and false-negative regions were concentrated near low-contrast boundaries, reflective surfaces, narrow furniture legs, and areas where the floor texture was visually similar to wall or cabinet surfaces. These errors are important because boundary noise can shift the estimated floor polygon and consequently affect the homography transformation.

Figure 5. Fine-tuned model prediction: (a) Original image, (b) Ground Truth mask, (c) Model prediction, (d) Overlay

The improvement after local fine-tuning is consistent with the importance of scene-specific training reported by Sun et al. [6]. Direct numerical ranking against [7] and [8] is avoided because their datasets, class definitions, and evaluation protocols differ from those used in this study. Therefore, the most defensible baseline in this manuscript is the stage-wise comparison between the pre-trained model and the locally fine-tuned model under the same architecture and measurement protocol.

d. Perspective Transformation Results

The module estimated four correspondence points from the mask contour and used cv2.findHomography to transform tile patterns under different room conditions. The warped rows became progressively smaller toward the far floor boundary, producing visually coherent depth under partial furniture occlusion. MAE compared the four predicted and manually annotated interior angles in two representative room images covering conditions with and without excessive furniture occlusion. Table 5 reports angular deviations only and therefore does not fully represent visible corner position shifts.

Table 5. Four Corner Angular Error of Perspective Estimation

Viewpoint Scenario	TL Error (°)	TR Error (°)	BL Error (°)	BR Error (°)	MAE (°)
With Furniture	8.35°	8.70°	8.61°	8.44°	8.53°
Without Excessive Furniture	5.95°	6.60°	0.04°	0.68°	3.32°
Average MAE					5.92°

TL, TR, BL, and BR are the four corners used to estimate the homography matrix. The metric reports interior-angle differences and does not measure pixel displacement; therefore, visible positional shifts may still occur even when angular MAE is small. Average MAE was calculated from the full-precision per-condition values before rounding to two decimal places.

The room without excessive furniture obstruction produced an MAE of 3.32°, whereas the room with furniture occluding part of the floor produced an MAE of 8.53°. The average MAE across both conditions was 5.92°. This two-image comparison indicates that furniture occlusion can increase perspective-estimation error, although a larger stratified test set is needed to confirm the effect of furniture density.

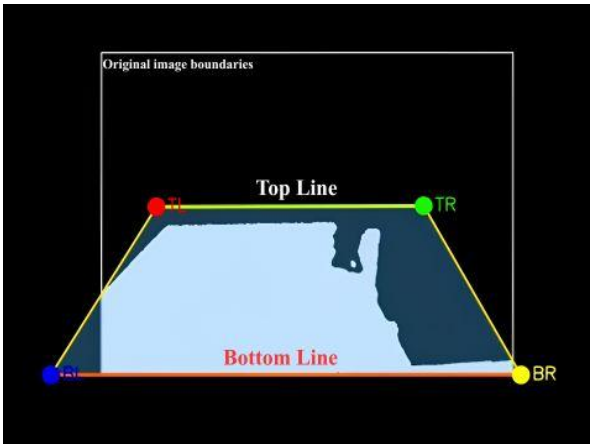

Figure 6. Homography correspondence point estimation

Figure 7. Ceramic tile motif mapped onto floor via perspective transformation

e. System Implementation Results

The web-based VTO system was successfully implemented with two main interface pages. The first page allows users to upload room photos and select ceramic tile motifs from the

catalog. The second page displays the VTO visualization results with the selected ceramic motif mapped onto the detected floor area. The system supports all 50 Granite Tile motif variations from the PT Concord Industry catalog.

f. Landing Page

This page presents an interactive interface where users can upload interior room images and select various Granite Tile motifs from the digital catalog. Users can view the available products and upload images independently, making the selection process more efficient, as shown in the image below.

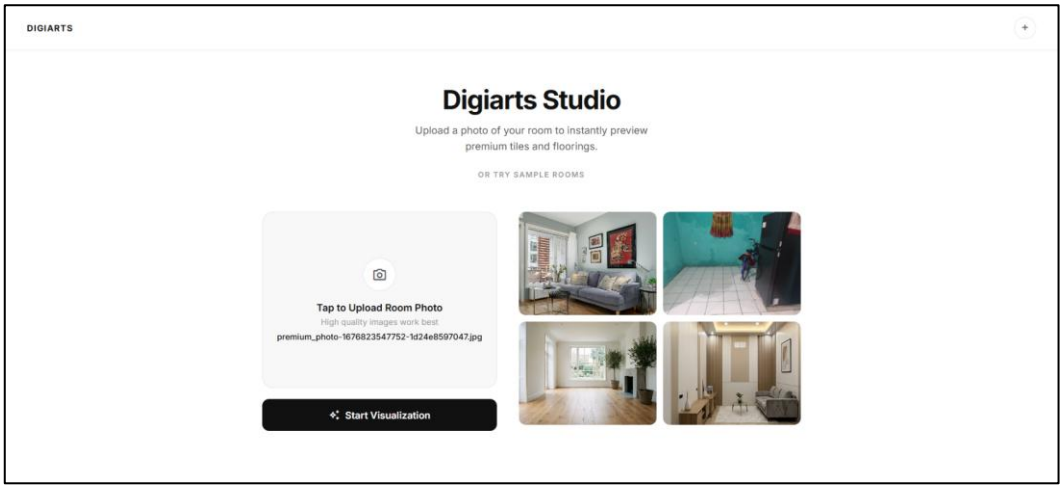

Figure 8. System interface room photo upload page

g. Virtual Try-On Visualization Page

This page provides a centralized interface for users to see the result of the perspective transformation. Users can instantly view how the selected ceramic motifs look realistically mapped onto their floor area through a single dashboard, making the visualization process highly accurate, as shown in the image below.

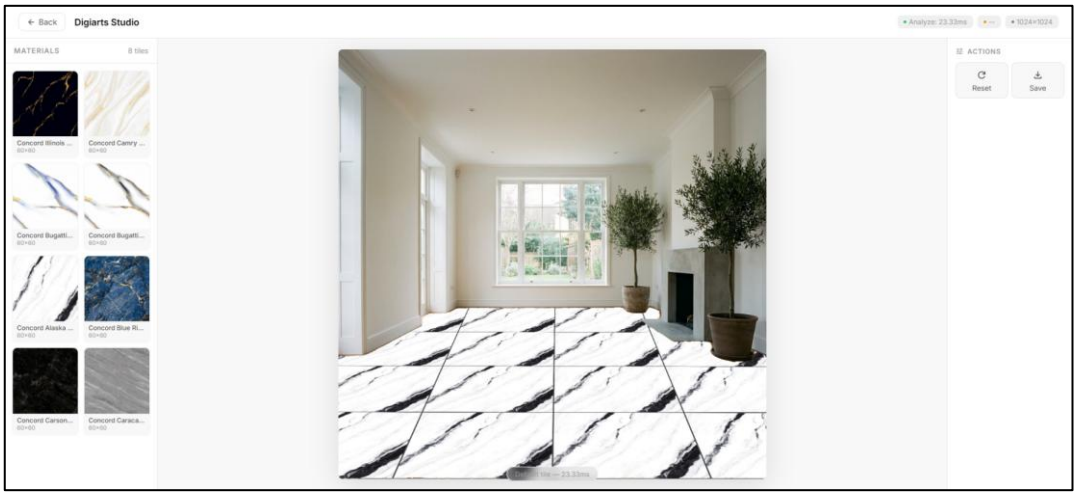

Figure 9. Virtual Try-On result page with ceramic motif selection panel

h. System Testing Results

Functional testing was conducted using the Black Box Testing method to validate all system features against design specifications.

Table 6. System Testing Results

No	Module	Number of Scenarios	Test Coverage	Status
1	Image Upload	3	Valid format, invalid format, oversized file	Valid
2	Floor Segmentation	4	Standard room, complex room, low light, partial occlusion	Valid
3	Motif Selection	2	Motif catalog display, motif selection	Valid
4	Perspective Transformation	3	Oblique angle, frontal angle, complex geometry	Valid
5	Visualization Output	2	Result display, image quality validation	Valid

All 14 functional scenarios produced the expected outputs. Unlike the marker-based application in [5], the proposed workflow does not require a physical marker. These black-box results verify system functionality, whereas consumer usability and purchase-decision effects require a separate respondent-based study.

5. Conclusions

This study developed a markerless floor tile VTO system by integrating U-Net segmentation and perspective transformation. The two-stage training strategy produced a test-set IoU of 0.9483, while homography mapped tile motifs according to the estimated floor geometry. Across conditions with and without excessive furniture, the perspective evaluation achieved an average four-corner angular MAE of 5.92°, supporting the angular component of the alignment evaluation.

The scientific contribution is an integrated automatic floor-segmentation and markerless transformation workflow; its practical contribution is a web interface that lets users visualize tile motifs in their own room photographs. The system therefore demonstrates potential to reduce visual uncertainty, but its direct effect on consumer decisions has not yet been established through a formal user study.

The perspective evaluation was limited to two representative images, no separate encoder architecture was trained as a baseline, and respondent-based usability testing was outside the study scope. The perspective evaluation also measured angular deviation only and did not quantify pixel-level corner displacement. The ConvNeXt-V2-Huge encoder also requires substantial computation. Future work should evaluate more diverse rooms, compare lightweight encoder baselines, add corner-position error metrics, conduct standardized user testing such as SUS, measure real-time inference, and apply model pruning or knowledge distillation.

References

[1] K. Wahyu Sukayasa, “Kajian Struktur Bentuk Dan Ruang Pada Karya Komposisi 2 Dimensi Sebagai Media Peningkatan Pemahaman Estetika,” *CandraRupa : Journal of Art, Design, and Media*, vol. 3, pp. 90–96, 2024, doi: 10.37802/candrarupa.v3i2.799.

[2] Kementerian Perindustrian Republik Indonesia, “Industri Keramik Semakin Berdaya Saing Didukung Penerapan SNI Wajib,” 2025. [Online]. Available: https://bskji.kemenperin.go.id/2025/05/19/industri-keramik-semakin-berdaya-saing-didukung-penerapan-sni-wajib/

[3] C. Chen, J. Ni, and P. Zhang, “Virtual Try-on Systems in Fashion Consumption: A Systematic Review,” *Applied Sciences*, vol. 14, p. 11839, 2024, doi: 10.3390/app142411839.

[4] J. Chen, B. He, H. Zhu, and J. Wu, “The Implicit Preference Evaluation for the Ceramic Tiles With Different Visual Features: Evidence From an Event-Related Potential Study,” *Front. Psychol.*, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1139687.

[5] B. Basri, M. Sarjan, and M. A. Riadi, "Aplikasi Augmented Reality Dengan Pengujian Standar Iso 25010 Pada Media Promosi Penjualan Keramik Lantai Berbasis Android," *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 12, 2024, doi: 10.23960/jitet.v12i3.4689.

[6] R. Sun *et al.*, "Training Indoor and Scene-Specific Semantic Segmentation Models to Assist Blind and Low Vision Users in Activities of Daily Living," *IEEE Open J. Eng. Med. Biol.*, vol. 6, pp. 533–539, 2025, doi: 10.1109/ojemb.2025.3607816.

[7] H. T. Ngoc, N. N. Vinh, N. Q. P. Le, N. N. N. Nguyen, T. A. N. Le, and V. N. Dinh, "Enhancing Semantic Scene Segmentation for Indoor Autonomous Systems Using Advanced Attention-Supported Improved UNet," 2024, doi: 10.21203/rs.3.rs-4587262/v1.

[8] C. N. Phu, H. N. Huu, K. Le Minh, and H. T. Ngoc, "Enhanced Image Segmentation for Indoor Autonomous Systems Using RealSense Camera and U-Net With Advanced Encoder and Attention Mechanism," *Proceedings of the 2025 10th International Conference on Intelligent Information Technology*, pp. 67–72, 2025, doi: 10.1145/3731763.3731788.

[9] Y. Luo *et al.*, "A Review of Homography Estimation: Advances and Challenges," *Electronics (Basel).*, vol. 12, no. 24, p. 4977, 2023, doi: 10.3390/electronics12244977.

[10] T. Betsas, A. Georgopoulos, A. Doulamis, and P. Grussenmeyer, "Deep Learning on 3D Semantic Segmentation: A Detailed Review," *Remote Sens. (Basel).*, vol. 17, no. 2, p. 298, 2025, doi: 10.3390/rs17020298.

[11] A. Amer, T. Lambrou, and X. Ye, "MDA-Unet: A Multi-Scale Dilated Attention U-Net for Medical Image Segmentation," *Applied Sciences*, vol. 12, p. 3676, 2022, doi: 10.3390/app12073676.

[12] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, doi: 10.1109/access.2021.3086020.

[13] K. Toida, N. Kato, O. Segawa, T. Nakamura, and K. Hotta, "Homography VAE: Automatic Bird's Eye View Image Reconstruction From Multi-Perspective Views," *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Application*, pp. 626–632, 2025, doi: 10.5220/0013244800003912.

[14] J. Kim and W. Heo, "Interior Design With Consumers' Perception About Art, Brand Image, and Sustainability," *Sustainability*, vol. 13, no. 8, p. 4557, 2021, doi: 10.3390/su13084557.

[15] P. Kán, A. Kurtic, M. Radwan, and J. M. L. Rodríguez, "Automatic Interior Design in Augmented Reality Based on Hierarchical Tree of Procedural Rules," *Electronics (Basel).*, vol. 10, no. 3, p. 245, 2021, doi: 10.3390/electronics10030245.

[16] H. Turkmen and D. Akgun, "A Hybrid UNet With Attention and a Perceptual Loss Function for Monocular Depth Estimation," *Mathematics*, vol. 13, no. 16, p. 2567, 2025, doi: 10.3390/math13162567.

[17] Tedyyana, Agus, et al. "Threat modeling in application security planning citizen service complaints." *Indonesian Journal of Electrical Engineering and Computer Science* 28.2 (2022): 1020.