HoughSAM: Geometry-Guided Zero-Shot Segmentation of Main Incision in Cataract Surgery Using SAM2 for Skill Assessment

Weina Dai, Jay Paranjape, Shameema Sikder, Vishal Patel, and Swaroop Vedula*

Johns Hopkins University, Baltimore MD 21218, USA
swaroop@jhu.edu

Abstract. Accurate estimation of wound dimensions during the main incision step of cataract surgery is a critical indicator of surgical proficiency, since a properly made wound that allows the widest part of the incision knife to pass through minimizes the risk of bleeding. In this work, we propose a novel, training-free framework that integrates Segment Anything Model 2 (SAM2) with automatic prompt generation via the Hough Transform to segment incision trajectories and estimate wound dimensions directly from surgical videos. By leveraging geometric priors to guide zero-shot segmentation, our approach eliminates the need for task-specific segmentation annotations or model finetuning. We evaluated the method on the public Cataract-1K dataset, using expert surgeon annotations as ground truth. Quantitative results demonstrate high balanced accuracy in predicting incision architecture and size ratings as indicators of skill, using only the features derived from the predicted segmentation masks. In general, our findings highlight the potential of combining zero-shot generalizable segmentation models with classical geometric priors for fine-grained surgical video analysis and skill assessment.

Keywords: Wound Segmentation · Cataract surgery · Skill Assessment.

1 Introduction

Assessment of surgical skill is essential for training surgeons. Video-based assessment of surgical skill using machine learning can enable applications such as personalized feedback and surgical coaching [18,1]. To be useful for downstream applications, skill assessment must be granular and context-specific. However, the conventional approach is to assess skill globally over the duration of the procedure or activities within the procedure. In current literature, Kim et al. [11] developed deep learning (DL) methods to evaluate the global skill for the capsulorhexis step of cataract surgery. Similarly, Hira et al. [9] train a ResNet network incorporating temporal modules to classify capsulorhexis videos as performed by an expert or a novice. On the other hand, Gong et al. [7] developed a

* Corresponding author

network to evaluate the generalizability of global skill assessment across different hospitals. Global assessment does not explain specific details of how the procedure, or a part of it, was performed by the surgeon. The need for interpretable surgical AI systems has been increasingly emphasized for clinical adoption and feedback-driven training [17,10].

While recent methods such as CatSkill have introduced phase-wise cataract skill assessment and derived more interpretable metrics, many of these metrics remain indirect and video-tied, such as the ability to maintain centration of the eye within the operating microscope's field of view (LCFC)[6]. These developments motivate a further shift toward interpretable, *context-specific* evaluation of surgical maneuvers and clinically meaningful geometric characteristics. We illustrate our work with the main incision, which is a critical step in cataract surgery. Surgeons do this incision to construct the main wound through which they access the inside of the eye in all subsequent steps in the procedure. In addition, surgeons construct the wound to be self-sealing for optimal healing and protection against infection. Improper construction of the main wound can adversely impact multiple other steps, wound healing and eventually, patient outcomes. Therefore, skill assessment in term of the geometry of the wound provides the context specificity needed for feedback and coaching surgeons.

Assessment of the geometry of the main wound can be enabled by accurate segmentation of the video images. Previous research on segmentation of cataract surgery videos was focused on instruments and anatomical structures, but not for fine-grained tasks like segmenting and tracking incisions. The main incision in cataract surgery represents a transient tool–tissue interaction region rather than a persistent object, and its geometry is highly irregular and dynamic across frames. In addition, variability in lighting conditions, microscope viewpoint, instrument types, and patient anatomy further complicates segmentation. Creating sufficiently large datasets annotated on a pixel level for this task is labor-intensive and prone to inter-annotator variability. These challenges motivate the development of approaches that are less dependent on task-specific annotations and extensive supervised training.

Foundation segmentation models, such as the Segment Anything Model (SAM) [12] and its successor SAM2 [15], are promising approaches to segment fine-grained details like the main incision. These models act as promptable "universal" segmentation engines that can generate high-quality masks from simple spatial prompts (e.g., points, boxes, or masks) without finetuning. However, relying solely on manual prompts is impractical in a surgical video context, and naive prompting strategies often fail to capture the aforementioned transient interaction regions like the main incision. To address this limitation, we propose a hybrid framework that integrates classical geometric priors for automatic prompt generation. Since the limbus and pupil exhibit approximately circular geometry under surgical microscopy, we employ the Hough Transform to detect candidate circular contours in each frame. Prompt points sampled within these detected regions guide SAM2 to automatically segment the limbus and pupil. Notably, due to SAM2's sensitivity to occlusions, regions disrupted by the inci-

sion knife are excluded from the predicted masks. We then apply convex shape completion and mask subtraction to isolate the disrupted region and derive the wound area. From this predicted wound mask, we extract several features that quantify geometry of the main wound.

To our knowledge, there are no publicly available cataract surgery datasets with ground truth manual annotations of segmentation masks for the main wound. These annotations require intensive manual effort. Hence, to measure the effectiveness of our method, we asked an expert surgeon to evaluate the wound architecture and size via manual inspection. We predict the expert evaluation using the features on geometry of the main wound. The following are the main contributions of this paper:

1. A novel geometry-guided, training-free pipeline called **HoughSAM** that combines Hough-based automatic prompt generation with SAM2 for segmenting the main wound in cataract surgery videos.
2. Qualitative comparison with manual prompting of SAM2 and MedSAM, for wound segmentation.
3. A quantitative framework showing how segmentation-derived metrics of wound geometry can be used for objective surgical skill assessment. **The code and data annotations will be released.**

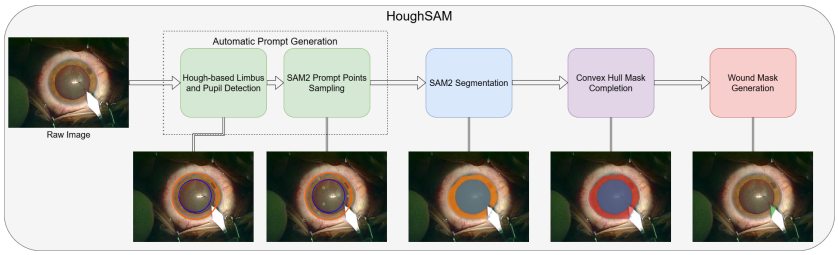

Fig. 1. HoughSAM pipeline

2 Methodology

2.1 Preliminaries

Definition of Main Incision Step in Cataract Surgery: In one technique, surgeons use a diamond-shaped knife called the keratome for the main incision. They insert the keratome through the cornea near the limbus in one plane and enter the anterior chamber in a different plane such that the wound seals by itself when the knife is removed. The main wound is completely formed when the widest part of the keratome passes the distal opening into the anterior chamber. A well constructed main wound with appropriate architecture:

1. is parallel to the iris,
2. is of appropriate size (width) to allow passage of surgical instruments, and
3. is self-sealing with minimal leakage, bleeding, and low risk of postoperative infection.

SAM2 and Its Limitations for Wound Segmentation: SAM2 is a promptable, foundation-scale segmentation model that generalizes across domains without task-specific finetuning. It consists of an image encoder that generates dense visual embeddings, a prompt encoder that incorporates user-provided spatial cues (e.g., points, bounding boxes, or masks), and a mask decoder that predicts segmentation masks conditioned on both image features and prompts. For video applications, SAM2 introduces temporal memory and prompt propagation mechanisms that enable mask tracking across frames. Given an initial prompt on a reference frame, the model can maintain spatial coherence and propagate segmentation results over time. This design makes SAM2 particularly effective for tracking well-defined, persistent objects across video sequences.

However, wound segmentation presents several challenges that limit the direct application of SAM2:

1. Transient and interaction-based nature – From a purely visual analysis perspective, the wound represents a temporary disruption of tissue rather than a stable semantic object, making wound segmentation fundamentally different from segmenting organs or tools. Its shape changes rapidly and may disappear once the blade exits the cornea.
2. No labeled datasets currently exist to support finetuning SAM2 or training specialized segmentation networks for this task, to the best of our knowledge.
3. Prompt sensitivity – In preliminary experiments, manually providing generic point or box prompts to SAM2 did not consistently yield accurate wound masks. This is likely because the wound does not correspond to a distinct object category in the model's training distribution.

Together, these limitations highlight that wound segmentation cannot be reliably achieved through naive prompting or object tracking alone. Instead, they motivate the development of a training-free, geometry-guided strategy that leverages stable anatomical structures to infer the transient wound region indirectly.

2.2 HoughSAM

Pupil and Limbus Mask Generation: Given a cataract video denoting the main incision step, we extract frames at a rate of 50 frames per second (fps). Each frame is first converted to grayscale, and a structured forest-based edge-detection algorithm [3] is applied to highlight the boundaries of the eye compartments and the incision edges. We then apply a standard Hough transform [4] to detect circles likely corresponding to the pupil and the limbus. Among all detected circles, we retain only the $N = 64$ closest to the center of the frame, based on our observation that camera recordings of cataract surgery usually focus on the limbus. Further, the limbus and pupil in a frame can be visually thought of as

circles with very close centres. Motivated by this, we find all valid pairs of circles $(c_1, r_1), (c_2, r_2)$ whose centers are close according to the constraint: $(abs(c_2 - c_1) < 0.1{*}r_1)$ and whose radius difference d falls within the range $[0.2{*}r_1, 0.5{*}r_1]$. Here, c denotes the centre and r denotes the radius of the circle. r_1 denotes the radius of the smaller circle (pupil candidate), and r_2 denotes the radius of the larger circle (limbus candidate). Likewise, c_1 denotes the centre of the smaller circle, and c_2 denotes the center of the larger circle. We select, among all valid pairs, the pair with the smallest radius difference as the circles corresponding to the pupil and the limbus. Finally, three equally spaced points that fall onto the circle $(c_1, r_1 - \frac{d}{2})$ are selected as point prompts corresponding to the pupil. Similarly, three equally spaced points that fall onto the circle $(c_1, r_1 + \frac{d}{2})$ are selected as point prompts corresponding to the limbus.

While Hough circles provide a coarse representation of the pupil and limbus, we require more anatomically consistent masks for wound detection. Hence, we employ SAM2 to segment them using the previously derived point prompts. We leverage SAM2's temporal memory mechanism to propagate masks across subsequent frames, maintaining spatial coherence. Thus, the resulting mask for frame f is denoted by $\mathbb{M}_L^f$ for the limbus and $\mathbb{M}_P^f$ for the pupil. This process is outlined with examples in Figure 1.

Preprocessing Module: It is important to note that since SAM2 is sensitive to occlusions, regions that are covered by the incision knife are typically excluded from these masks by SAM2. While this behavior may reduce anatomical completeness, it is advantageous for isolating the wound region. While major occlusion from the knife is required for wound segmentation, lighting variations and segmentation artifacts can also create minor holes and spurious regions in the mask. Thus, we apply morphological closing using OpenCV [2] to remove these holes and fill minor boundary discontinuities. Similarly, connected component filtering from is used to remove small spurious regions within the mask. In addition, we observed that SAM2 produces masks with jagged edges. Hence, for both $\mathbb{M}_L^f$ and $\mathbb{M}_P^f$, we perform polygonal approximation to remove such edges while preserving overall geometry. We show an example of this stage in Figure 1

Wound Segmentation: After the preprocessing step, the refined masks for a given frame may be unoccluded due to the absence of a surgical knife, or may contain the occlusion caused by the instrument. This occluded part contains the wound that lies in the interaction region between the knife and the limbus. To extract this, we first compute the convex hull $\mathbb{M}_{LC}^f$ from $\mathbb{M}_L^f$, which represents the smallest convex region enclosing the detected contour [8]. Next, we compute the raw mask $\mathbb{M}_K^f$ for the region of the knife inside the limbus as follows:

$$\mathbb{M}_K^f = \mathbb{M}_{LC}^f - \mathbb{M}_L^f \tag{1}$$

Finally, we remove the pixels belonging to the pupil region from this mask to form the final wound segmentation mask $\mathbb{M}_W^f$ as follows:

$$\mathbb{M}_W^f = \mathbb{M}_K^f \cap \neg\mathbb{M}_P^f \tag{2}$$

where $\neg\mathbb{M}_P^f$ denotes the logical inverse of the pupil mask. Thus, the final incision mask $\mathbb{M}_W^f$ highlights pixels that: (i) lie along the optic-disc boundary, (ii) correspond to concavities introduced by the blade, and (iii) exclude the interior of the pupil. This geometry-driven derivation enables robust localization of the incision region without requiring explicit wound annotations or task-specific model training, as shown in Figure 1.

3 Experiments

Data and Annotations: We used 10-second clips of 80 videos from the publicly available cataract-1k [5], which show the complete main incision step. From each clip, we selected frames at 50 fps, capturing incision initiation, propagation, and completion.

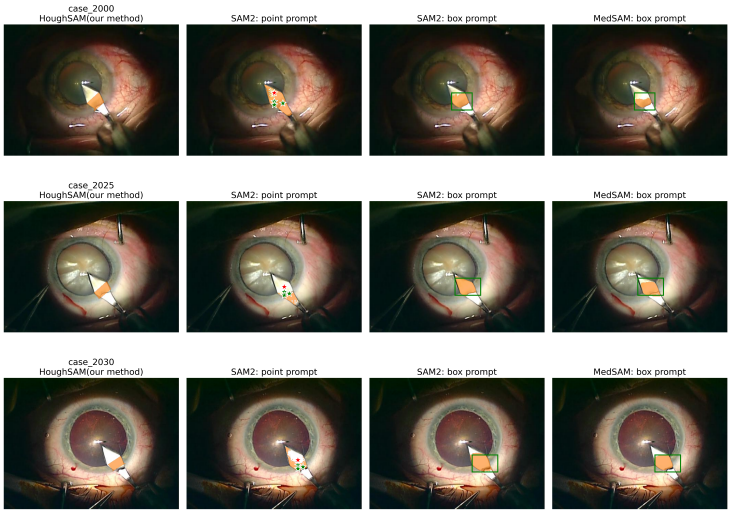

Fig. 2. Qualitative comparision between HoughSAM(our method) and direct prompting with SAM2 [15] and MedSAM [13]. Please zoom in for details.

3.1 Qualitative Results on Wound Segmentation

Figure 2 presents qualitative comparisons between our method and direct prompting strategies using SAM2(point prompts) [15], SAM2 (box prompts), and MedSAM (box prompts) [13]. Due to the absence of publicly available pixel-level wound annotations, no established supervised baselines exist for this task. To provide meaningful comparisons, we therefore construct reference baselines by

applying SAM2 and MedSAM with their preferred prompting strategies to directly segment the suspected wound region. As shown in the figure, direct prompting methods frequently either undersegment the incision region, leak into surrounding corneal tissue, or capture portions of the surgical instrument rather than the wound itself. In contrast, our geometry-guided approach consistently isolates the incision region without any manual prompting. Notably, by leveraging relatively stable anatomical structures (pupil and limbus) rather than attempting to segment the transient wound directly, our method achieves more stable and anatomically coherent results across diverse cases and lighting conditions. These qualitative results demonstrate that integrating classical geometric priors with foundation segmentation models enables reliable wound localization without task-specific training data.

3.2 Applications to Skill Assessment

To evaluate whether the information acquired from wound segmentation is useful, we asked an expert to review the main incision video segments and to rate the surgeon's skill on two criteria - incision architecture and size. We hypothesized that classifiers trained on the features extracted from the wound geometry can accurately predict these metrics. For this purpose, we aggregated frame-level wound masks into case-level geometric descriptors. Because incision quality is reflected not only in the instantaneous wound shape but also in its temporal evolution, we designed features that capture both spatial geometry and temporal dynamics. Specifically, we extracted the following features:

- **max_ratio**: the maximum ratio between the left and right sides of the wound. This measures whether the incision reaches sufficient and proportional width on both sides.
- **min_ratio**: the minimum ratio observed during the incision process, capturing early instability or incomplete knife passage.
- **max_deviation**: the maximum deviation of the wound's side ratio from the ideal ratio of 1. This reflects misalignment or poor entry angle.
- **early_area_growth_rate** and **late_area_growth_rate**: the average slope of the wound area curve during the early phase and late phase of knife insertion. We define the early phase to be the period between wound initiation and reaching the maximum intersection area, and define the late phase to be the period between maximum intersection area and complete knife removal. A smooth, controlled incision exhibits gradual area increase, whereas abrupt or unstable insertions produce irregular patterns.
- **max_area**: the maximum wound area achieved during the incision step, corresponding to passage of the widest part of the knife.
- **min_area**: the minimum wound area.

Experimental Results: To evaluate whether wound-derived features are predictive of surgeon skill, we trained lightweight classifiers at the case level using the above features. Given the limited dataset size (80 segments), we adopted

shallow decision trees and small multilayer perceptrons with cross-validation for model selection. For comparison, we used pretrained foundation models CLIP [14] and CSMAE [16] features for the segment of the video covering the incision phase and passed those to the trainable classifier. CLIP is pretrained with natural images while CSMAE is trained on specifically cataract video data. The dataset was split into training, validation, and test subsets at case level with (0.6 : 0.2 : 0.2) ratio, and hyperparameters were selected via grid search. For both **incision architecture** and **incision size**, the best test balanced accuracy, 0.580 and 0.625 respectively, were achieved with decision tree classifiers on the held-out test set, as shown in Table 1. This is likely due to class imbalance, as the majority (around 70%) of the cases in our dataset are labeled as "good". Nevertheless, the results demonstrate that wound-derived geometric features contain meaningful predictive signals and produce much better results than generic CLIP features. In addition, HoughSAM features are also able to outperform CSMAE features [16] that were specifically trained for cataract videos, showing the effectiveness of extracting task-specific features for downstream tasks.

Table 1. Quantitative results for incision quality. Acc denotes the balanced test accuracy for each metric

Input Features	Model	Incision Architecture Acc	Incision Size Acc
CLIP features	Decision Tree	0.333	0.301
CSMAE features	Decision Tree	0.233	0.429
HoughSAM features	Decision Tree	**0.580**	**0.625**
HoughSAM features	MLP	0.579	0.495

4 Limitations

Our framework has several limitations. First, multiple stages of the pipeline rely on heuristic geometric constraints and empirically chosen thresholds. While these heuristics were motivated by typical anatomical priors for cataracts, they may compromise robustness under challenging imaging conditions. Second, the proposed method assumes that the limbus and pupil appear approximately circular and are viewed near frontally, which is generally true in cataract surgery videos but may not hold under extreme viewpoint changes or severe occlusions. Third, the expert skill labels were provided by a single annotator, which may introduce subjective bias in the evaluation. Future work will explore adaptive geometric modeling, multi-rater validation, and learning-based prompt generation to improve robustness across diverse surgical settings.

5 Conclusion

In this work, we presented a training-free framework for automatic segmentation of the main incision and extraction of context-specific features on the

geometry of the main wound in cataract surgery videos. By integrating Hough-based geometric prompt generation with SAM2, and convex hull–based shape reasoning, we demonstrated that transient incision regions can be reliably derived without manual prompting or task-specific finetuning. This work highlights the potential of combining foundation models with classical geometric priors for fine-grained surgical video analysis in data-limited settings. At the same time, several limitations remain. The dataset size for skill prediction is modest, and class imbalance affects balanced accuracy, particularly for minority ratings. In addition, our method assumes approximately circular limbal geometry and may be sensitive to extreme viewpoint changes or severe occlusions. Future work will explore larger-scale validation, more robust temporal modeling, and integration with multimodal surgical signals to further strengthen objective skill assessment.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmidi, N., Poddar, P., Jones, J.D., Vedula, S.S., Ishii, L., Hager, G.D., Ishii, M.: Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. International Journal of Computer Assisted Radiology and Surgery **10**(6), 981–991 (Jun 2015). https://doi.org/10.1007/s11548-015-1194-1, https://doi.org/10.1007/s11548-015-1194-1
2. Bradski, G.: The opencv library. In: Dr. Dobb's Journal of Software Tools (2000)
3. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1841–1848 (2013). https://doi.org/10.1109/ICCV.2013.231
4. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Communications of the ACM **15**(1), 11–15 (1972). https://doi.org/10.1145/361237.361242
5. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k: Cataract surgery dataset for scene segmentation, phase recognition, and irregularity detection (to appear)
6. Giap, B.D., Ballouz, D., Srinivasan, K., et al.: Catskill: Artificial intelligence-based metrics for the assessment of surgical skill level from intraoperative cataract surgery video recordings. Ophthalmology Science **5**(4), 100764 (2025). https://doi.org/10.1016/j.xops.2025.100764
7. Gong, Z., Wan, B., Paranjape, J.N., Sikder, S., Patel, V.M., Vedula, S.S.: Evaluating the generalizability of video-based assessment of intraoperative surgical skill in capsulorhexis. International Journal of Computer Assisted Radiology and Surgery **20**(11), 2231–2239 (Nov 2025). https://doi.org/10.1007/s11548-025-03406-0, https://doi.org/10.1007/s11548-025-03406-0
8. Graham, R.L.: An efficient algorithm for determining the convex hull of a finite planar set. Information Processing Letters **1**(4), 132–133 (1972)
9. Hira, S., Singh, D., Kim, T.S., Gupta, S., Hager, G., Sikder, S., Vedula, S.S.: Video-based assessment of intraoperative surgical skill. International Journal of Computer Assisted Radiology and Surgery **17**(10), 1801–1811 (Oct 2022). https://doi.org/10.1007/s11548-022-02681-5, https://doi.org/10.1007/s11548-022-02681-5

10. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
11. Kim, T.S., O'Brien, M., Zafar, S., Hager, G.D., Sikder, S., Vedula, S.S.: Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. International Journal of Computer Assisted Radiology and Surgery **14**(6), 1097–1105 (2019). https://doi.org/10.1007/s11548-019-01956-8, https://doi.org/10.1007/s11548-019-01956-8
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023). https://doi.org/10.48550/arXiv.2304.12306, https://arxiv.org/abs/2304.12306
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
15. Ravi, N., Mao, H., Rolland, C., Gustafson, L., Mintun, E., Xiao, T., Whitehead, S., Berg, A.C., Dollár, P., Girshick, R., Kirillov, A.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024). https://doi.org/10.48550/arXiv.2408.00714, https://arxiv.org/abs/2408.00714
16. Shah, N.A., Bandara, C., Sikder, S., Vedula, S.S., Patel, V.M.: Csmae : Cataract surgical masked autoencoder (mae) based pre-training. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2025). https://doi.org/10.1109/ISBI60581.2025.10981288
17. Tonekaboni, S.M., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: Contextualizing explainable machine learning for clinical end use. Machine Learning for Healthcare Conference (2019)
18. Vedula, S.S., Ishii, M., Hager, G.D.: Objective assessment of surgical technical skill and competency in the operating room. Annual Review of Biomedical Engineering **19**, 301–325 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044435, https://doi.org/10.1146/annurev-bioeng-071516-044435