ScanTailor Spectre

Version 2.0a16 | macOS (Apple Silicon) | Requires macOS 12 or later

ScanTailor Spectre transforms raw scans into clean, publication-ready pages. Import a PDF or folder of images, process through a 9-stage workflow, and export a polished, searchable PDF.

ScanTailor Spectre is not intended for copyrighted works, but rather for works that you have the rights to or are in the public domain.

What's New in Spectre

Version 2.0a16

- **PDF File Association** - Open PDFs directly from Finder via "Open With" or drag-and-drop
- **Improved OCR Text Alignment** - Text highlights now accurately match visible characters in exported PDFs

Version 2.0a15

- **Auto Levels** - One-click tonal correction for cleaner pages
- **PDF DPI Detection** - Import honors embedded DPI for more accurate sizing
- **Image Quality Improvements** - Better results across output processing
- **Apply To Fixes** - Despeckle and Depth Perception batch dialogs behave correctly
- **Output Performance** - Faster illumination normalization, transforms, and TIFF writing

Version 2.0a14

- **Portable Projects** - Projects now use relative paths; move projects between machines without relinking
- **Large Project Export** - Fixed PDF export for books with 100+ pages
- **Performance Boost** - Parallel Sauvola binarization, vImage SIMD acceleration, Metal GPU optimizations

Features

- **Apple Silicon Native** - Gatekeeper-friendly application
- **PDF Import** - Open PDFs directly, no need to extract pages first
- **Batch Processing Summaries** - Dialogs after stages 2, 4, 5 to catch problems and jump to pages needing attention
- **Detection Settings** - Adjustable Fill Factor and Border Tolerance for art books and photo-heavy content
- **Finalize Stage** - New stage for color mode selection and output format
- **Intelligent Color Detection** - Auto-detects B&W vs grayscale vs color pages, including embedded photographs
- **White Balance Tools** - Auto white balance and pick-paper-color for correcting aged paper and lighting
- **OCR Stage** - Automatic text recognition for searchable PDFs
- **Export Stage** - New dedicated PDF export with quality presets

---

Quick Start

1. Import Your Scans

From the startup screen or from the file menu: - **Import PDF** - Extract pages from an existing PDF - **Import Folder** - Load a folder of scanned images - **Import Project** - Load a project you have previously saved.

Supported formats: PDF, TIFF, PNG, JPEG, BMP

Recently saved projects will be shown here.

2. Work Through the Filters

Each filter stage refines your pages. Work top to bottom:

#	Stage	Purpose
1	**Fix Orientation**	Rotate pages right-side up
2	**Split Pages**	Separate two-page spreads
3	**Deskew**	Straighten tilted scans
4	**Select Content**	Crop to content area
5	**Margins**	Set page margins
6	**Finalize**	Choose B&W, Grayscale, or Color
7	**Output**	Apply image processing
8	**OCR**	Text recognition for searchable PDFs
9	**Export**	Create final PDF

To run a stage, click on the little triangular play button next to the stage you are at. Stage 9 is special. You don't need to run it, you simply click on Export to PDF.

3. Batch Process

You may want to try doing everything automatically if your scan is clean. If you need to split spreads to pages, run stage 2. If not, or if you have run stage 2 already go to stage 6, Finalize, and run it and each of the next stages.

After batch processing certain stages, a summary dialog appears to help catch common problems:

- **Stage 2 (Split Pages)** - Shows pages that were/weren't split, lets you force changes
- **Stage 4 (Select Content)** - Shows pages with low content coverage that might need layout preserved
- **Stage 5 (Margins)** - Warns about unsplit spreads or outlier pages that cause margin issues. If you have unspilt spreads, you will have to go back to Stage 2.

The dialogs in stage 2 and 5 let you jump directly to problem pages and take corrective action without hunting through hundreds of thumbnails.

The Workflow Explained

Each stage has automatic operations and potentially requires minimal input, but some scans can be more difficult to process than others. For example, a decent scan with only text pages might require virtually no input on your part. A 500 page book with color pages, some of which are skewed, many of which have color casts and require individual adjustments may require substantial intervention.

Although ScanTailor Spectre is designed to be as automated as possible, it can make mistakes. It's always best to check your output.

After you run a stage, the parameters that stage set are stored in your project on a per-page basis. Let's say you set a page to color once. If you rerun the automatic color detection, it will not alter that setting even if it thinks it is grayscale.

This is pre-beta software. Save your project frequently.

Stage 1: Fix Orientation

Corrects pages that are upside-down or rotated. Auto-detection handles most cases; use the rotate buttons for manual override. It is rare that you will have to even run this separately unless your scans are awful. This is not the same as Stage 3, Deskew.

Stage 2: Split Pages

Separates book spreads (two pages per scan) into individual pages. The split line is auto-detected but can be adjusted manually.

After batch processing, a summary dialog appears with two toggle buttons: - **Not Split (X)** - Pages kept as single pages (may need to be split) - **Split (Y)** - Pages that were split (may have been split incorrectly)

Double-click any page to jump to it for inspection. Use the action buttons to force changes: - In "Not Split" view: **Force Two-Page** makes pages split - In "Split" view: **Force Single Page** makes pages unsplit

Visual indicators in the list: - Gray strikethrough = page was visited/jumped to - Dark green strikethrough = action was taken (force split or unsplit)

Stage 3: Deskew

Straightens pages that were slightly tilted during scanning. Even 1-2° of tilt is noticeable in output.

Tip: Sort by "decreasing deviation" to review the most-skewed pages first. The sort order is located at the bottom of the thumbnail panel on the right.

Stage 4: Select Content

This stage has two separate but related functions: defining the **Page Box** (the physical page boundaries) and the **Content Box** (the area containing actual content like text and images).

When do you need this? The Page Box helps crop out artifacts on the sides of the image like margins or the bed of the scanner. The Content Box identifies where text or images are so margins can be standardized. For most workflows, Auto mode handles both well, but you may need Manual adjustments for pages with unusual layouts (title pages, illustrations, fold-outs) or when auto-detection picks up shadows or artifacts as content.

Page Box

The Page Box defines the boundaries of the physical page within the scan. This is useful when: - The scanner captured more than just the page (scanner bed edges, background) - You want to define a consistent page size across all scans - The automatic page detection didn't work correctly

Page Box options: - **Disable** - Don't detect page boundaries; use the entire image - **Auto** - Automatically detect the page edges (looks for the transition from scanner background to paper) - **Manual** - Draw the page boundary yourself by dragging corners and edges

When Auto mode is selected, **Fine Tune Page Corners** adjusts corner positions by looking for black edges, useful for books where corners may be shadowed or bent.

In Manual mode, you can enter exact **Width** and **Height** values to set a specific page size.

Content Box

The Content Box defines what part of the page contains actual content. Everything outside this box becomes white margin in the final output.

Content Box options: - **Disable** - Don't detect content; use the entire page - **Auto** - Automatically find text and images on the page - **Manual** - Draw the content area yourself

Detection Settings

When using Auto mode for Content Box, these settings control how content is detected:

- **Fill Factor** (0.50-1.00, default 0.65) - Controls what density of content is accepted. The default is optimized for text documents. For art books, photo-heavy pages, or pages with large dark areas, increase this value:
  - **0.65** - Default, works well for text documents
  - **0.85** - Better for mixed text/image content
  - **0.95+** - Use the entire page as content (best for art books, covers, full-page images)
- **Border Tolerance** (0-20px, default 2) - Controls how much content touching page edges is preserved:
  - **Low (0-5)** - Aggressively trim edge content (removes shadows, scanner artifacts)

- **High (15-20)** - Preserve content that extends to page edges (for images that bleed to margins)

Tip: For art books or pages with large photographs, set Fill Factor to 0.95+ to capture the entire page content. The default settings are optimized for text documents and may clip images.

This excludes: - Scanner bed edges - Book margins you want to remove - Fingers or page holders

Auto detection is not always perfect, especially with images or decorative elements. Sort by "Order by completeness" to review pages that may need manual adjustment.

Content Coverage Summary

After batch processing, a summary dialog appears showing pages grouped by content coverage: - **Low Coverage** - Pages where detected content covers less than the threshold percentage of the page area. These may have intentional layout (chapter titles, short paragraphs) that shouldn't be centered. - **Normal Coverage** - Pages where content fills most of the page area.

Use the **Coverage threshold** slider (20%-80%) to adjust what counts as "low coverage."

Actions for low coverage pages: - **Jump to Selected** - Navigate to a page for manual inspection - **Preserve Layout (Selected/All)** - Disable content detection for these pages, keeping their original layout intact

Visual indicators: - Gray strikethrough = page was visited/jumped to - Dark green strikethrough = action was taken (preserve layout)

Stage 5: Margins

Sets white space around content in the final output and page size.

- **Top/Bottom/Left/Right** - Individual margin sizes
- **Alignment** - Where content sits within the page

Stage 6: Finalize

This stage determines how each page will be processed:

- **Color Mode**: Black & White, Grayscale, or Color
- **Output Format**: TIFF or PNG (you can ignore this if you don't care to save the individual images)
- **Output Location**: Where processed files are saved

The app auto-detects the appropriate color mode. Obviously, Color pages require the most storage space, Grayscale requires less, and Black and White the least of all.

The **Midtone Threshold** slider adjusts detection sensitivity, useful for smaller images (lower = more pages will be inspected and detection will be slower). You can override the determination after inspection.

White Balance (Color and Grayscale modes): - **Force white balance** - Automatically corrects color casts from aged paper or poor lighting - **Pick Paper Color** - Click this, then click on an area that should be white/neutral in the preview. The app will correct the entire page based on that sample.

Clear All Pages resets all pages to unprocessed state, useful if you want to re-run automatic detection with different settings.

Previous versions of ScanTailor exported to images, the workflow in this one is images or PDF to PDF. Images will be discarded when you close the project unless you choose to preserve them at this point by ticking the box to "Preserve Output Images."

Stage 7: Output

Applies image processing to generate the output files. Options vary by color mode.

Output Resolution: Sets the DPI of the output image. 400 DPI is the Library of Congress recommendation for most documents. Higher = sharper but larger files.

Black & White Mode Options

- **Binarization Threshold** - Controls the cutoff between black and white. Lower values = more black, higher = more white. Adjust if text appears too thin or too bold.
- **Normalize Illumination** - Evens out lighting variations across the page. Helps with scans that have shadows or uneven exposure.
- **Morphological Smoothing** - Smooths jagged edges on text and lines.
- **Morphological Opening/Closing** - Advanced cleanup options for removing small artifacts.

Grayscale & Color Mode Options

- **Normalize Illumination** - Evens out lighting variations.
- **Sharpen/Blur filters** - Enhance or soften the image.

Brightness & Contrast

- **Brightness** - Adjusts overall image lightness. Slide right to brighten, left to darken.
- **Contrast** - Adjusts tonal range. Slide right to increase contrast, left to decrease.

The sliders have tick marks at -100, 0, and +100, with a center detent that snaps to 0 for easy reset.

Paper Detection (Equalize Illumination)

When "Equalize illumination (Color)" is enabled, these advanced controls appear:

- **Min Brightness** (0-255, default 120) - Pixels brighter than this may be paper
- **Max Saturation** (0-255, default 60) - Pixels less saturated than this may be paper
- **Min Coverage** (0-50%, default 1%) - Minimum paper area required for equalization
- **Adaptive (sample margins)** - Auto-detect paper color from page margins

These controls help the algorithm distinguish paper from content when correcting uneven lighting.

Common Options (All Modes)

- **Dewarping** - Flattens curved pages from book spines. Set to "Auto" for automatic detection or "Manual" to adjust control points yourself. Most useful for thick books where pages curve near the spine.
- **Despeckle** - Removes small dots and noise:
  - *Cautious* - Minimal cleanup, preserves detail
  - *Normal* - Balanced approach
  - *Aggressive* - Heavy cleanup, may remove fine details
- **Equalize Illumination** - Additional lighting correction.
- **White Margins** - Ensures margins are pure white.

Zones

- **Picture Zones** - Draw rectangles around photographs or illustrations. These areas are processed differently to preserve detail and gradients instead of being binarized.
- **Fill Zones** - Draw areas to fill with a solid color (usually white). Useful for removing stamps, stains, or unwanted marks.

To draw a zone, select the zone tool, then click and drag on the preview image.

Apply To…

Once you've tuned settings for one page, use **Apply To…** to copy those settings to other pages. You can apply to: - All pages - Selected pages only - Pages with the same color mode

This is essential for efficiently processing large documents.

Stage 8: OCR

Performs optical character recognition to make your PDF searchable. When enabled, text is recognized and embedded as an invisible layer in the exported PDF, allowing you to search and select text.

Options: - **Enable OCR** - Toggle text recognition on/off (enabled by default) - **Language** - Select the document language for better recognition accuracy, or leave on Auto-detect - **Accurate Recognition** - Use accurate mode for best results (slower) or fast mode for quicker processing

Status Display: - Shows whether the current page has been processed - Displays the number of text blocks found

OCR runs automatically when you batch process through the Export stage. Results are cached per-page, so re-running won't re-process pages that already have OCR data unless you clear them.

Note: OCR uses Apple's Vision framework and requires macOS 13.0 or later.

Stage 9: Export

Creates the final PDF.

Max DPI: Limits output resolution for grayscale and color pages.

400 DPI is the Library of Congress recommendation for most documents. If your scan is lower quality or to be read on screen only, you may find that lowering the resolution as low as 72 DPI gives acceptable results and a much lower final PDF size. At 72 DPI, however, it would be very difficult to read black and white pages, which is why they are controlled by the master output resolution in Stage 7.

Keyboard Shortcuts

Key	Action
`Page Up/Down`	Previous/next page
`Home` / `End`	First/last page
`Cmd+S`	Save project

Tips

Scanning

- **300-400 DPI** for text documents
- **600 DPI** for fine detail or small text
- Scan in **color** even for B&W content - better conversion results

Processing

- Work through stages **in order** - each depends on the previous
- Configure **one page well**, then batch apply to similar pages
- Use **page ordering** options to find problem pages
- **Save frequently** - all settings are preserved in the project file

Color Mode Selection

- **B&W (1-bit)**: Smallest files, pure black and white, ideal for text-only
- **Grayscale**: Better for pages with photos or illustrations
- **Color**: Only when color information matters

Typical File Sizes (100-page book)

- B&W PDF: 5-15 MB
- Grayscale PDF: 20-50 MB
- Color PDF: 50-150 MB

Project Files

Projects are saved as `.ScanTailor` files containing: - References to source images (not copies) - All settings for every stage - Processing state

Important: Keep source images in place - projects reference them by path.

Credits

ScanTailor Spectre is based on: - ScanTailor Advanced by 4lex4 - ScanTailor by Joseph Artsimovich

License

GPL-3.0 - See LICENSE for details.

ScanTailor What?

ScanTailor Spectre is a Mac-native fork of ScanTailor Advanced, developed by Claude. Several considerations informed the choice of its name.

First, the name serves to distinguish it from other community releases, including ScanTailor, ScanTailor Advanced, and ScanTailor Experimental.

Second, the title references the opening line of the Communist Manifesto: “A spectre is haunting Europe.” Jacques Derrida, in “Spectres of Marx” wrote that the dead refuse to remain absent. For Derrida, this refers to the spectre of Marxism, but it can also refer to the haunting of the Internet by the ongoing persistance of ScanTailor as well as to the scanned book existing as a ghostly trace circulating online rather than as a physical object.

Fourth, ScanTailor Spectre is developed using Claude Code. The name also draws attention to the role of unauthorized book scans in the development of large language model artificial intelligences. It is illegal to copy and distribute copyrighted books, and we do not condone using ScanTailor Spectre for such purposes. Both Anthropic and Meta trained their models on millions of books obtained from shadow libraries such as Library Genesis, which are motivated by a vision of information freedom championed by Aaron Swartz. Recently, Anthropic settled with authors for $1.5 billion, equating to approximately $3,000 per book, a sum that exceeds the lifetime earnings of most books. In contemporary society, intellectual property is controlled not by individuals, but by those in positions of power.

Lastly, the first application I vibe-coded was a clone of the 1990s game Spectre VR.

The “spectre in the machine” draws on all of these.