UniMERNet: A Universal Network for Real-World Mathematical Expression Recognition

Supplementary Material

Contents

A. UniMER Dataset Details

A.1. Dataset Collection

SPE and CPE Sampling Existing datasets, such as IM2LATEX-100k and Pix2tex, present two primary challenges. Firstly, the size of these datasets, typically ranging from 100k to 200k formulas, is insufficient for training a precise and robust MER model. Secondly, these datasets contain a limited number of complex formulas, which compromises the model's performance, particularly in handling multi-line complex expressions.

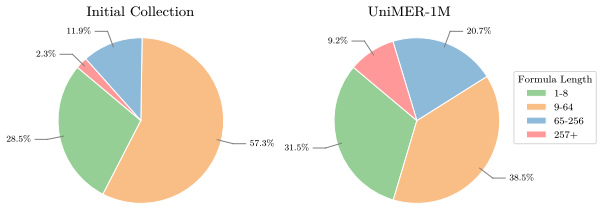

Figure 1. Formula length before and after re-sampling.

To address the limited size of the dataset, we expand it by incorporating an additional 4 million LaTeX expression source codes, building on the previously mentioned open-source datasets. These new entries are predominantly sourced from arXiv (89%), with supplementary contributions from Wikipedia (9%) and StackExchange (2%). This initial dataset expansion enhances the model's overall capabilities. However, the proportion of long formulas in the initial collection is relatively small (2.3%), which may cause inadequate training for complex expressions. To address this, we extract the longest formulas as CPE and adjust their ratio with randomly sampled SPE. This rearrangement ensures a balanced representation of varying lengths within the dataset, thereby significantly improving the model's ability to recognize multi-line mathematical expressions. The distribution after rearrangement is shown in Figure 1.

SCE Deduplication When extracting mathematical formulas from PDF pages, we face a unique challenge: formulas originating from the same page often appear identical in content, leading to potential duplicates. Simple deduplication based on textual content alone risks significant data loss, as identical formulas can appear across different pages, each bearing distinct visual characteristics such as font styles, sizes, and backgrounds. To preserve the richness of visual diversity while eliminating true duplicates, we adopted an image-based deduplication strategy, employing Perceptual Hashing to assess image similarity. This method allows us to compare the visual features of the formula images directly, ensuring that only those with high similarity—indicating true duplicates—are removed. Through this meticulous process of image similarity analysis, we effectively reduced the dataset to 4,744 unique Screen-Captured Expressions (SCE), each representing a distinct visual instance of mathematical expressions, thereby constituting our refined SCE test set. This curated collection provides a more reliable benchmark for evaluating model robustness, ensuring that performance measurements reflect genuine generalization capabilities across diverse visual renderings encountered in real-world mathematical documents.

Rendering Settings For the rendering settings, we follow a similar procedure used in [2] and [1]. The dataset is rendered using XeLaTeX with a diverse range of math fonts and DPI settings. The chosen fonts included Asana Math, Cambria Math, XITS Math, GFS Neohellenic Math, TeX Gyre Bonum Math, TeX Gyre Dejavu Math, TeX Gyre Pagella Math, and Latin Modern Math, with Latin Modern Math as the default math font being employed in approximately 22% of the cases. To accommodate different levels of clarity and detail, the DPI setting varies between 80 to

350 when converting to PNG format, allowing for adjustments in the resolution and sharpness of the rendered mathematical expressions for improved dataset diversity.

A.2. Formula Text Normalization

LaTeX syntax inherently contains ambiguous information, as different source codes can produce the same rendering. This presents significant challenges in the evaluation phase of the math formula recognition task's benchmark because it potentially leads to incorrect assessments of a model's performance despite producing visually identical formula renderings. In handwritten math recognition datasets, such as CROHME, a self-defined label graphs format is used, eliminating ambiguous expressions by employing a character relation-based method. We do not adopt these methods for normalization as they involve format conversions during model training and, more importantly, only partial LaTeX syntax is supported with this approach.

The LaTeX normalization is first introduced in [2]. This preprocessing operation involves fixing super-script and sub-script order, replacing ambiguity with unified expressions while resulting in no or minimal visual changes in rendering, preserving the integrity of the original mathematical expressions. Subsequent datasets, such as IM2LATEX-100K [2] and Pix2tex [1], have adopted similar methods. Building on this foundation of normalization, we have adjusted the normalization rules for certain LaTeX environments, enabling better support for complex formulas and previously unsupported syntax. All formulas in UniMER-1M and UniMER-Test undergo this normalization process to facilitate better horizontal comparison with previous datasets that employ a similar normalization process.

A.3. Data Statistics

Most Occurring Symbols Diving into the dataset's LaTeX symbols offers intriguing insights into the most frequently utilized mathematical notations. The bar chart provided in the Figure 2 illustrates the frequency of specific LaTeX symbols that appear in UniMER. Symbols such as Greek letters, operators, and various mathematical functions are universally prevalent in each dataset, underlining their fundamental role in articulating complex mathematical ideas. A subtle variation is observed in the SCE and HWE datasets, where numbers and letters are noticeably more frequently occurring, as they contain relatively easier and less structured math expressions.

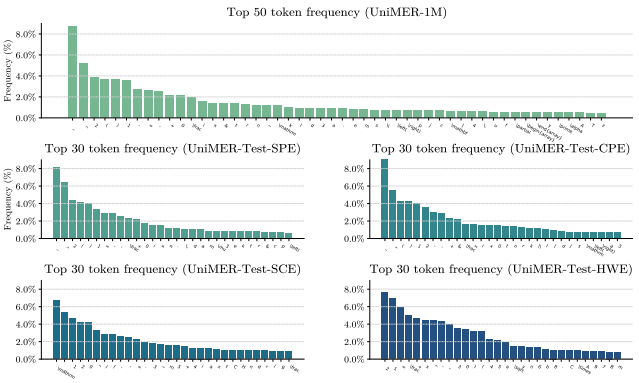

Figure 2. Most frequently occurring LaTeX symbols in UniMER-1M and UniMER-Test subsets

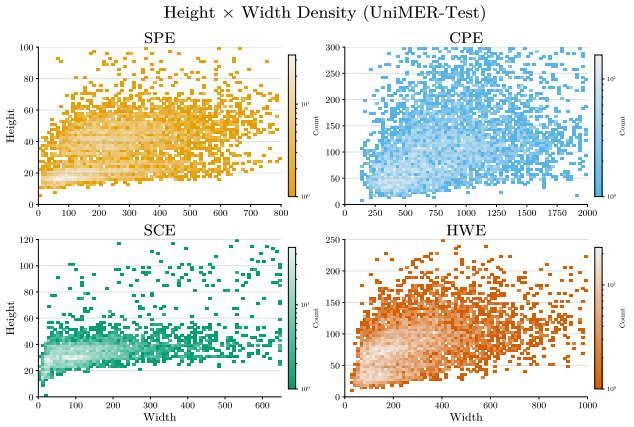

Figure 3. Height width density plot in UniMER-Test subsets

Image Size Distribution The scatter plot in Figure 3 provides a visual distribution of image sizes across different subsets within the UniMER-Test. Each point on the plot represents an individual image, with its position determined by the image's width and height. The SPE, CPE, SCE, and HWE subsets each exhibit unique clusters, indicating the variety in dimensionality they encompass. It's evident that the SPE and SCE subsets tend to have a higher concentration of smaller images, as shown by the dense clustering of points towards the lower end of the spectrum. The distribution of image sizes within the CPE dataset exhibits a considerable spread, highlighting the diversity of dimensions that this particular subset encompasses, indicating its complexity compared to SPE. On the other hand, the HWE subset is characterized by images with generally larger dimensions. This can be attributed to the fact that these images are often photographed and contain noise, necessitating a higher resolution to ensure that the finer details of the handwritten expressions are preserved and recognizable.

B. Additional Experiments

B.1. Evaluation on IM2LATEX-100K

To further contextualize our model's performance on established printed formula benchmarks, we evaluated UniMER-Net on the IM2LATEX-100K test set. This benchmark resembles the SPE subset of UniMER-Test, featuring

clean, computer-rendered formulas. As shown in Table 1, UniMERNet achieves state-of-the-art performance, outperforming previous methods on both CDM and BLEU metrics, demonstrating its robustness on standard benchmarks.

Table 1. Performance on the IM2LATEX-100K test set.

Method	CDM ↑	BLEU ↑
WYGIWYS	-	87.73
WAP	-	88.21
UniMERNet	**0.994**	**92.56**

B.2. Comparison with Fine-tuned LMMs

To create a fair comparison with Large Multimodal Models (LMMs), we fine-tuned a moderate-sized LMM, Qwen-2.5 VL-Instruct (3B), on our UniMER-1M dataset. As shown in Table 2, fine-tuning significantly boosts the LMM's performance, confirming the high quality and diversity of our dataset. Despite the LMM's 10x parameter count (3B vs 0.3B), UniMERNet equipped with R-S attention still surpasses it on the challenging CPE subset (0.972 vs 0.952) and achieves comparable or superior results on SPE and HWE, highlighting the efficiency and effectiveness of our task-specific architecture.

Table 2. Comparison with zero-shot and fine-tuned LMM (Qwen-3B) on UniMER-Test (CDM ↑).

Method	SPE	CPE	SCE	HWE
Qwen-3B (zero-shot)	0.970	0.837	0.935	0.900
Qwen-3B (fine-tuned)	0.986	0.952	**0.953**	0.950
UniMERNet (Window Attn)	0.991	0.931	0.939	0.929
UniMERNet (R-S Attn)	**0.991**	**0.972**	0.940	**0.954**

B.3. Evaluation on MathWriting Dataset

We also evaluated UniMERNet on the MathWriting dataset, using Character Error Rate (CER) as the metric. As detailed in Table 3, UniMERNet achieves a CER of 3.45, significantly surpassing other methods like PaLI (5.95) and CTC Transformer (5.49) by 2.5 and 2.04 points, respectively.

Table 3. Performance on the MathWriting dataset (CER ↓).

Method	MathWriting (CER ↓)
PaLI	5.95
CTC Transformer	5.49
UniMERNet	**3.45**

B.4. Encoder Backbone Ablation

We further study the choice of visual encoder backbone. The comparison keeps decoder and training settings fixed, varying only the visual encoder backbone. As shown in Table 4, Swin consistently outperforms ViT on challenging formula subsets while also reducing parameters and improving throughput.

Table 4. Encoder backbone ablation under the same training setup (CDM ↑).

Encoder	HWE	CPE	Params	FPS
ViT	0.931	0.935	564M	7.67
Swin-Transformer	**0.941**	**0.955**	**313M**	**10.48**

B.5. Qualitative Comparisons

As shown in Figure 5, we selected two representative samples from the UniMER-Test set to thoroughly compare the performance between UniMERNet and other SOTA methods. Notably, while the other models exhibit certain shortcomings in handling these test samples, our model consistently delivers robust and accurate recognition results, even with long and structurally complex expressions.

C. Training and Model Analysis

C.1. Training Time Data Augmentation

While introducing additional training data in UniMER-1M enhances the variety of formulas, it does not account for the diversity of real-world formula images, which can come from scanned documents or photos and can exhibit noise and distortion. We employ various image augmentation techniques during model training to simulate this diversity with extra transformations from Albumentations [1] library and self-defined transformations, which include but are not limited to:

- **Erosion/Dilation** - To simulate the textural imperfections often found in screen-captured formulas, these operations modulate the thickness of characters, mirroring the effects of resolution differences and printer anomalies.
- **Degradation Simulation** (Fog, Frost, Rain, Snow, Shadow) - These augmentations introduce environmental artifacts to mimic the conditions under which documents might be photographed in real-world scenarios, adding layers of complexity such as blurriness and occlusions.
- **Geometric Transformations** (Rotation, Distortion . . .) - To account for the angle and perspective distortions typical in photographed or scanned documents, these operations adjust the orientation and shape of the mathematical expressions.

[1] https://albumentations.ai

Each image undergoes a sequence of these augmentation operations with a given probability. This helps to bridge the gap between the pristine, synthesized training data and the noisy, real-world test images and improves UniMERNet's performance for real practical use. Figure 6 provides a visualization of selected transformations.

C.2. Training Loss with R-S Attention

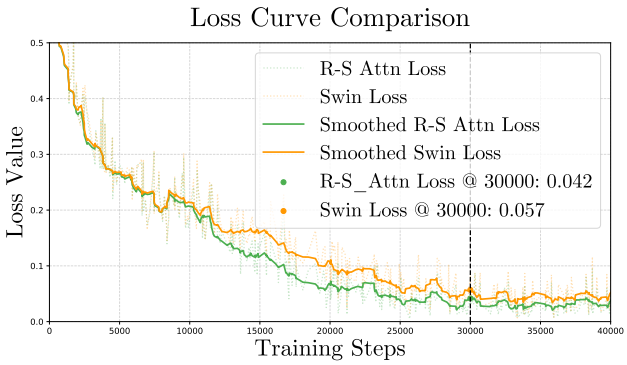

Figure 4. Training loss comparison of UniMERNet with different attention mechanisms over 40,000 iterations.

We plot the training loss of UniMERNet for 40,000 iterations and find that when using R-S Attention, the loss converged more quickly compared to other attention mechanisms. This is possibly due to the R-S Attention constraining how the model focuses on relevant regions, leading to more efficient learning and faster convergence. As shown in Figure 4, the training loss with R-S Attention decreases significantly faster than with Window Attention, demonstrating its effectiveness in optimizing the training process.

C.3. Expanded Ablation on H/V Processing Order

For completeness, we provide an expanded order ablation of Raster-Scan Attention. We reverse the standard processing flow from Horizontal → Vertical (H→V) to Vertical → Horizontal (V→H), while keeping all other settings unchanged. The results in Table 5 show that H→V outperforms V→H on CPE/SCE/HWE and ties on SPE, with the clearest gain on CPE (+0.007 CDM), further supporting the reading-order inductive bias hypothesis.

Table 5. Expanded order ablation of Raster-Scan Attention on UniMER-Test (CDM ↑).

Order	SPE	CPE	SCE	HWE	FPS
V → H	0.991	0.948	0.937	0.938	10.47
H → V (ours)	**0.991**	**0.955**	**0.939**	**0.941**	**10.48**

C.4. Robustness Analysis

In real-world applications, document images often suffer from rotational distortions due to improper scanning angles, handheld device tilts during capture, or perspective distortions in natural scene images. For mathematical expression recognition tasks, such geometric variations can severely impact symbol alignment perception and structural relationship modeling, which are critical for accurate parsing. An important question is whether the R-S attention method proposed, which relies on vertical and horizontal reading features, will be vulnerable to rotated input.

To evaluate the model's robustness under rotational perturbations, we created a rotated test set by applying random rotations between -10° to +10° to the original UniMER-Test dataset. This evaluation setting specifically avoids using rotation-based data augmentation during training to maintain the rigor of robustness assessment. The experimental results are presented in Table 6, where bold values indicate the best performance and underlined values denote the second-best results.

Key observations from the experimental results:

- Global attention demonstrates the strongest baseline robustness compared to window attention and pure R-S attention, with minimal performance degradation across all subsets.
- Pure R-S shows significant performance drops (up to 0.255 CDM loss in CPE subset), revealing its vulnerability to geometric transformations.
- The hybrid attention mechanism in UniMERNet effectively balances structural understanding and robustness, achieving comparable performance to global attention with better scalability.
- Input size scaling (+x2) combined with hybrid attention yields the most robust configuration, reducing performance loss by up to 40% compared to baseline methods.

These results validate our architectural design choices, particularly the effectiveness of combining window-based and R-S attentions for handling geometric variations in mathematical expression recognition tasks.

D. Limitations

While UniMERNet and its Raster-Scan Attention achieve strong performance and efficiency, they rely on the predominant sequential, raster-scan nature of mathematical expressions. Although our experiments (Sec. C.4) demonstrate that the hybrid attention mechanism and input scaling enhance robustness to rotational distortions, performance might still be challenged by highly unconventional, nonlinear layouts. Secondly, despite the extensive diversity of the UniMER-1M dataset, the model's ability to generalize to entirely novel mathematical symbols, notations, or extreme degradation types not represented in training remains

Table 6. Rotation robustness analysis of different attention mechanisms. Bold values indicate best performance, underlined values show the second-best performance.

Method	SPE			CPE			SCE			HWE		
	CDM ↑	Rot CDM ↑	Loss ↓	CDM ↑	Rot CDM ↑	Loss ↓	CDM ↑	Rot CDM ↑	Loss ↓	CDM ↑	Rot CDM ↑	Loss ↓
Global attn	0.991	0.833	<u>0.158</u>	0.943	0.749	<u>0.194</u>	0.935	0.729	0.206	0.939	0.893	<u>0.046</u>
Window attn	0.991	0.812	0.179	0.931	0.724	0.207	0.939	0.734	0.205	0.929	0.879	0.050
Only R-S attn	0.991	0.789	0.202	0.938	0.683	0.255	0.937	0.726	0.211	0.934	0.867	0.067
UniMERNet (Window+R-S)	0.991	0.807	0.184	0.955	0.722	0.233	0.939	0.738	<u>0.201</u>	0.941	0.881	0.060
UniMERNet (Window+R-S) †	**0.995**	**0.888**	**0.107**	**0.972**	**0.850**	**0.122**	**0.940**	**0.795**	**0.145**	**0.954**	**0.926**	**0.028**

†Model with 384×1344 input resolution ($2\times$ standard).

a potential constraint. Future work could explore enhancing robustness to a wider spectrum of structural variations and out-of-distribution inputs.

D.1. Failure-Case Quantitative Audit

To provide additional insight, we further audited sample-level changes introduced by R-S attention. Compared with a non-R-S baseline, R-S improves recognition on approximately 33% of samples while causing degradation on only about 1.5%. The rare regressions are concentrated in extreme long-form CPE cases (typically sequence length $>$ 1000), where adjacent matrix-like columns have weak visual anchors and may be partially merged. In practice, stacking multiple blocks and interleaving local window attention alleviates this issue, but these extreme layouts remain a key direction for future robustness improvements.

References

[1] Lukas Blecher. pix2tex - latex ocr. https://github.com/lukas-blecher/LaTeX-OCR, 2022. Accessed: 2024-2-29. 1, 2

[2] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. Image-to-markup generation with coarse-to-fine attention. pages 980–989. PMLR, 2017. 1, 2

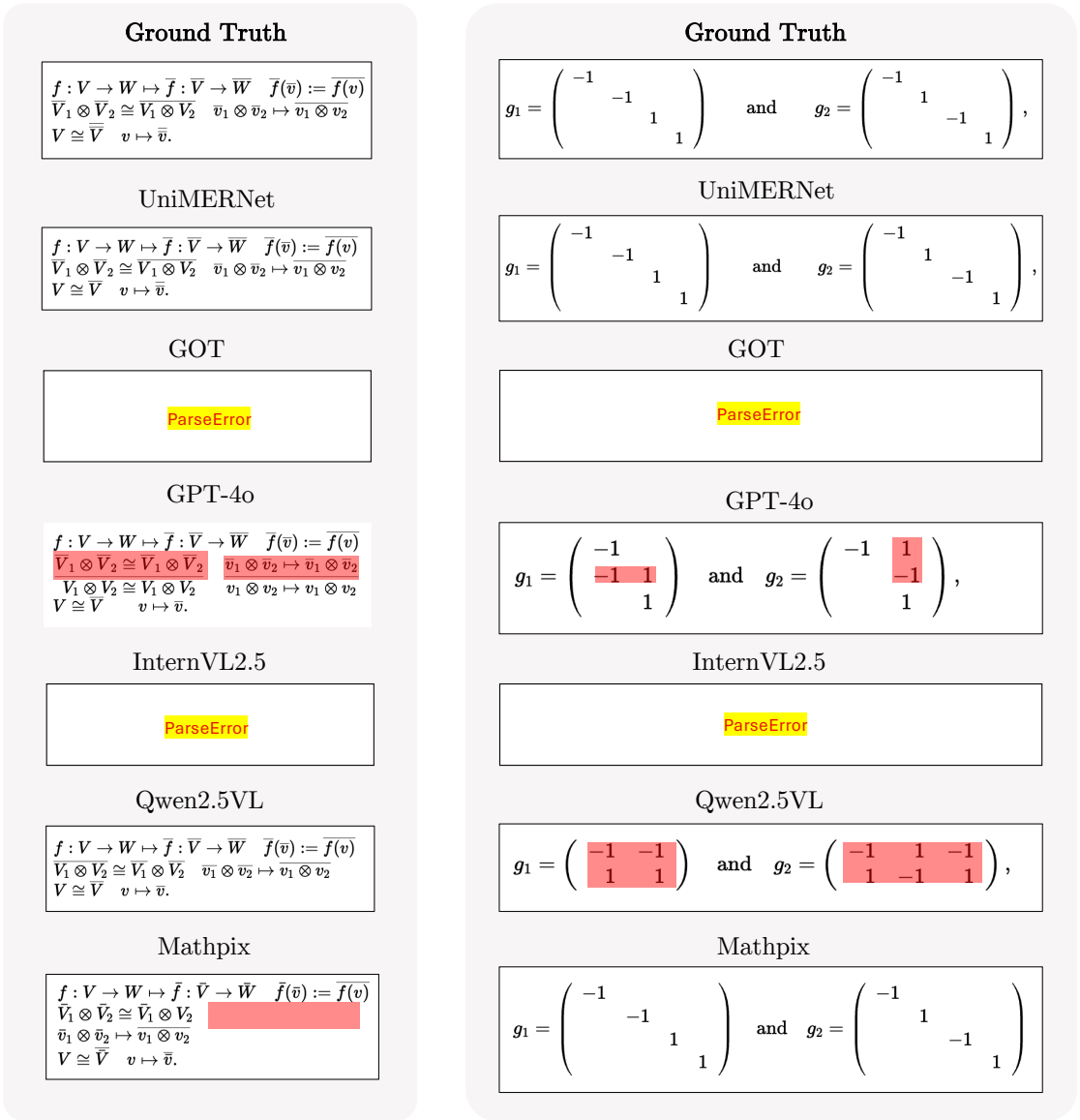

Figure 5. Visual comparison with other SOTA methods.

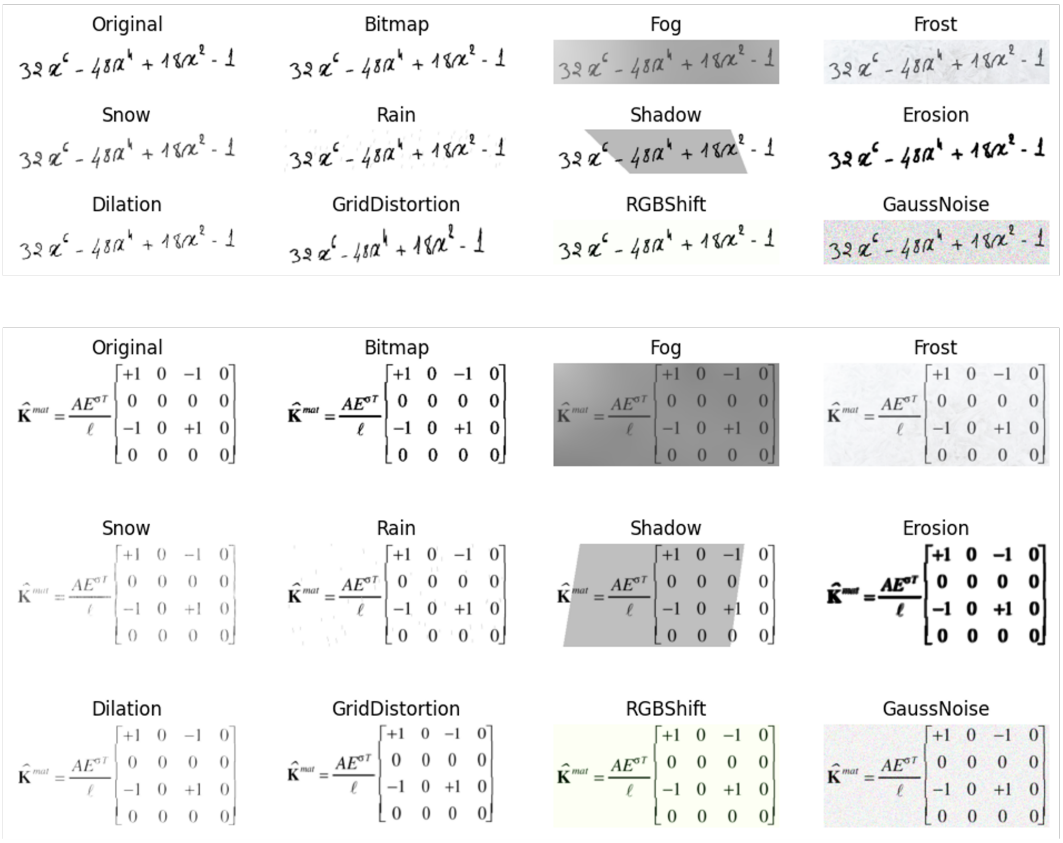

Figure 6. Visualization of selected image augmentations applied during training.