On the Effectiveness of Heterogeneous Ensemble Methods for the Re-identification of Industrial Entities - Supplementary Material

No Author Given

No Institute Given

1 Trustworthiness

1.1 Motivation

A largely unstudied benefit of ensembles is the increased trustworthiness they can provide to tasks like re-identification. Next to the increased performance, ensembles also tend to generalize better to unseen data [2] by allowing sub-model errors to cancel each other out [1].

Since we use heterogeneous sub-models, we can also use the difference in errors between sub-models to learn more about our algorithms and dataset. This can allow us to remove biases introduced by different sub-models to mitigate sensitivities, e.g., to changes in brightness or shearing effects. By using simple approaches to re-identification, it could also be extended to provide explanations for why two samples are considered similar to each other. Visualizing the sub-models results might provide an increase in explainability and trustworthiness: It allows the resulting matches and mismatches to be studied and to find and alter features that lead the models to make erroneous predictions. We believe this to be of great value to users, providing them with models that hold a greater degree of explainability and which can therefore ultimately be considered to be more trustworthy. Additionally an important part of trustworthiness is the reliability of the results over multiple experiments. Thus, considering it a measure of trustworthiness, we also study the uncertainty of our results, by employing cross-validation.

1.2 Results

Finally, we want to study the effect that using ensembles has on increasing the reproducibility, reliability and explainability of our resulting model. First, we present the relative uncertainty of the Rank-10 accuracy of each of our individual and ensemble approaches in Fig. 1.

We choose to present Rank-10 instead of Rank-1 accuracy, as these uncertainties are naturally quite noisy, which is reduced by the higher rank. Both ensembles that outperform each individual model in terms of accuracy also have the lowest uncertainty. They reach an uncertainty that is almost an order of magnitude smaller than the *average color* approach and clearly lower than all

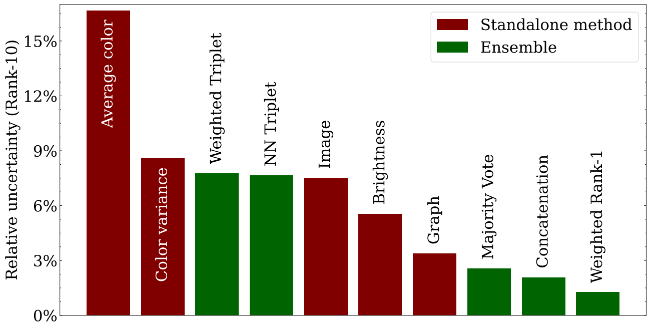

Fig. 1. Relative uncertainty (accuracy uncertainty divided by accuracy) of the Rank-10 accuracy for different models.

individual models. This increased reliability can be a crucial benefit to the implementation of our method in industrial settings.

Finally, we present a visualization of our sub-models' ranked results in Fig. 2. This visualization shows how different methods can come to different results, i.e., in this instance the comparison between image-based, graph-based, and color variance-based models that lead to different rankings of the same set of images, making it possible to debug methodological problems.

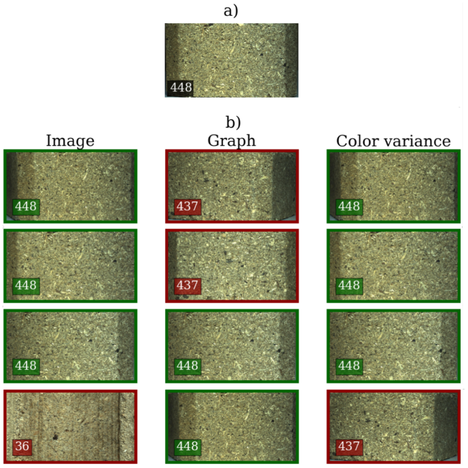

Fig. 2. Rank-1 to Rank-4 gallery matches (b) to a query image (a), for three sub-models. The Image ID is indicated in the corner, correct matches are framed in green, incorrect ones in red.

2 Additional Analysis

2.1 Time Efficiency

While each sub-model improves the ensemble performance, some of our models require significantly more time to be trained than others. This leads to a trade-off between training effort and re-identification performance. Motivated by this, we display both the Rank-1 accuracy and the training duration of partial ensembles in Fig. 3.

An ensemble of all three *color*-based methods can be trained in a few seconds and still performs almost as well as the *graph*-based method with a significantly higher training cost. This highlights another use case of heterogeneous ensembles: besides increasing the overall performance of a re-identification method it may also contribute to decreasing the time required for achieving comparable performance. We expect an ensemble of many more simple sub-models to outperform single complex models and be easier to use.

2.2 Contribution Studies

To study our ensembles further, we present the Rank-1 accuracy improvement of an ensemble compared to its best individual model for all two-model ensembles in Fig. 4.

First, it is apparent that almost all Rank-1 changes are positive. Only the combination of a *graph* representation with an *average color* leads to a negative change in Rank-1 accuracy. We hypothesize that this might result from the high

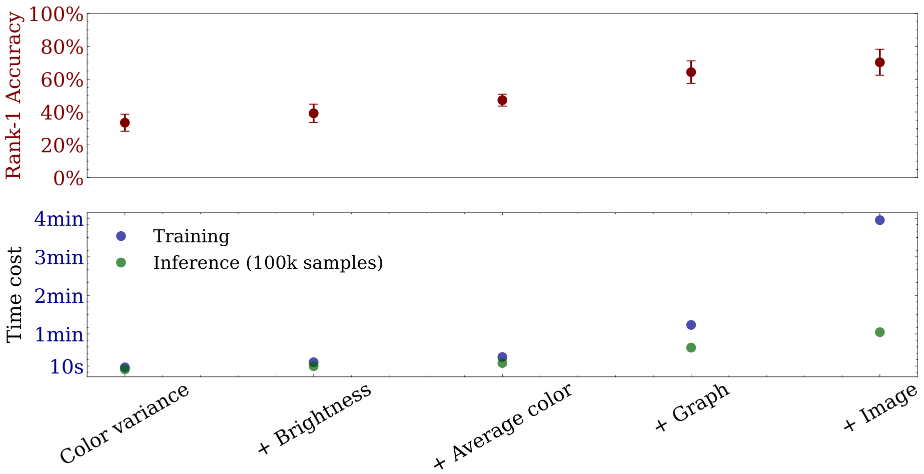

Fig. 3. Training duration, prediction cost and Rank-1 accuracy of different ensembles. The sub-models are chosen to minimize the training duration. All our experiments were conducted using an NVIDIA A100 graphics card with 40GB of VRAM.

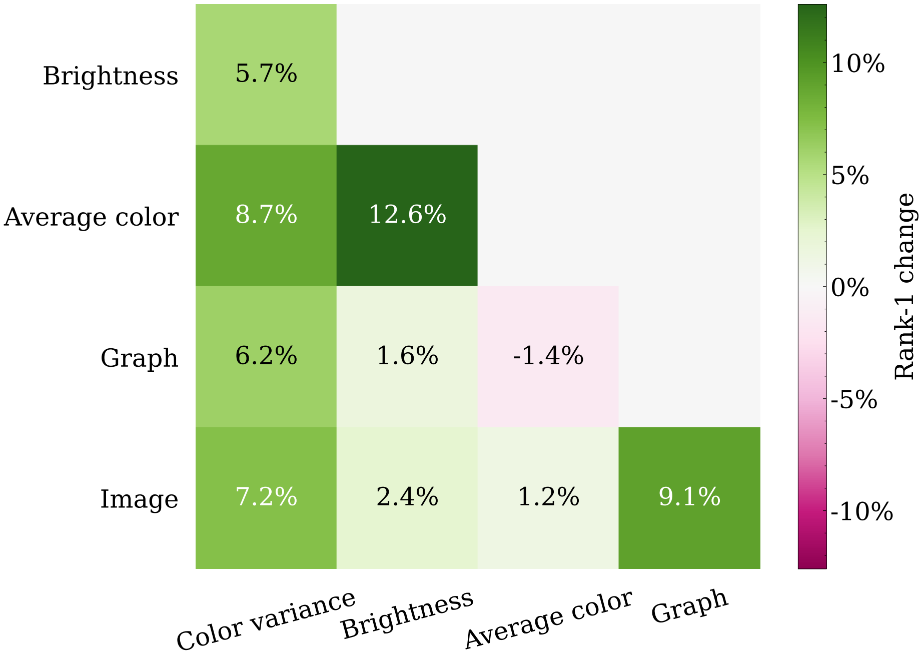

Fig. 4. Rank-1 accuracy change for all possible two-model ensembles compared to their best sub-model performance.

uncertainty of the *average color* model and the drastic difference in the performance of both models. Most importantly, the sum of each row and column is positive, implying that adding another model always benefits the ensemble (see also Fig. 2 in the supplementary material). This might also explain why we do not need any weighting factors to achieve higher performance; and it implies that even more ensemble sub-models might help to improve the ensemble's performance further.

2.3 Ablation study on representation sizes

We tested various representation sizes for the color variance approach, to determine which representation size to choose (in this case, a representation size of 50). This is shown in Fig. 5.

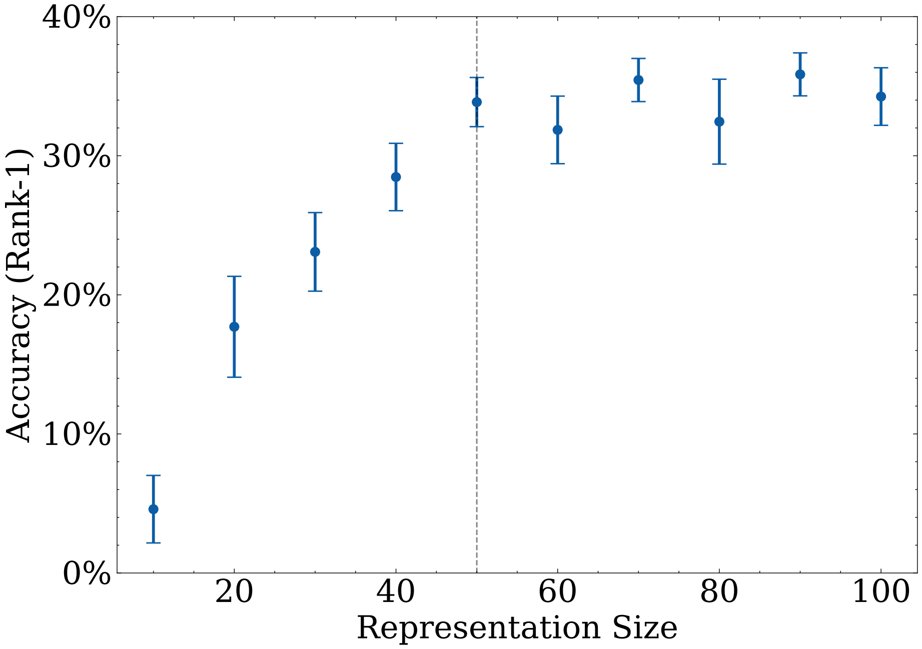

Fig. 5. Ablation study: Rank-1 accuracy of the color variance approach, as a function of the representation size used.

Here, the performance seems to plateau from a representation size of 50 onward, which is why we decided to use it. Similarly we chose the same size for the rest of the sub-models, except for the image-based network. Here, we chose a higher number (100), as the higher number of input features might relate to a higher number of output features.

2.4 Singular sub-model impact

While we study the contribution of individual sub-model pairs in the main paper (Fig. 4), another way to characterize these contributions would be to study ensembles with only 4 out of 5 sub-models included. Thus, removing the most important sub-model shows the lowest resulting performance. This is studied in Fig. 6.

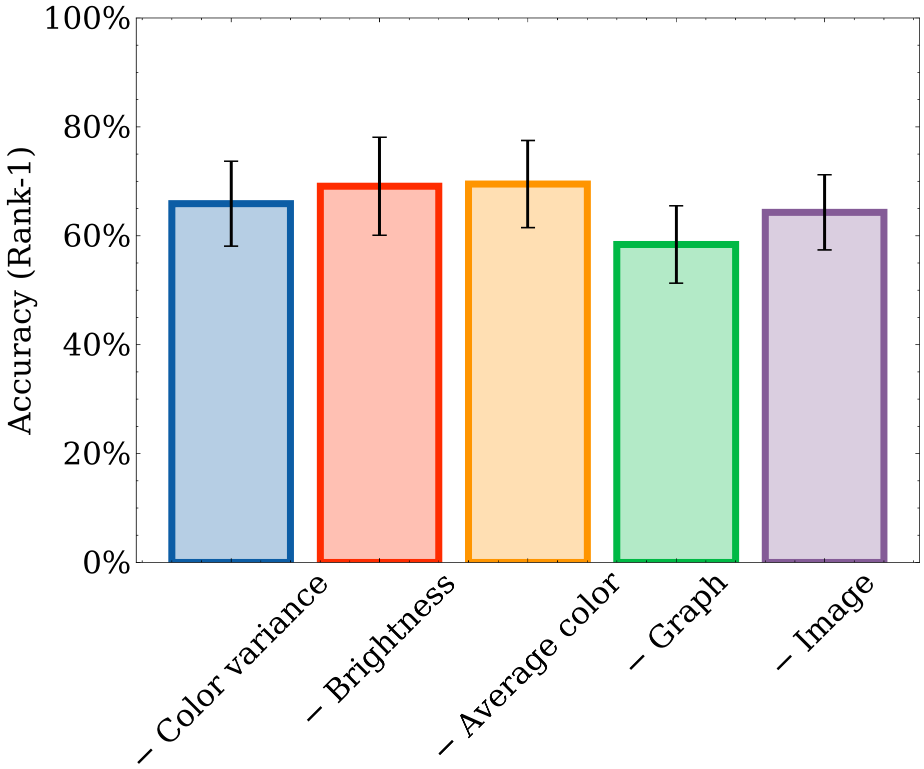

Fig. 6. Ablation study: Rank-1 accuracy for our ensemble approach when removing one of the sub-model (the respective one noted on the x-axis).

While the graph shows that, as expected, the highest impact is achieved by the graph model, the differences are not significant, showing that which specific ensemble sub-models to use might not be the most important choice to make.

2.5 Sub-model correlation

We want to characterize how far the predictions of different sub-models differ from one another. Because we cannot simply use the Pearson Correlation for this, as the sub-model predictions are vectors, we define triplet correlations in Alg. 1. The correlation between two functions is calculated via the likelihood that said functions provide the same order for the distances between three samples. The results are normalized so that the correlation follows the usual range of -1 to 1, with 0 representing a random chance.

Algorithm 1 Calculation of $corr_{triplet}$

```
Require: $f$,$g$, $x \in X$, $accuracy$
  $count \leftarrow 0$
  $success \leftarrow 0$
  while $count < accuracy$ do
      $sample\ (A, B, C)\ fromX$
      $\Delta_f^{A,B} = \|f(A) - f(B)\|_2$
      $\Delta_f^{A,C} = \|f(A) - f(C)\|_2$
      $\Delta_g^{A,B} = \|g(A) - g(B)\|_2$
      $\Delta_g^{A,C} = \|g(A) - g(C)\|_2$
      if $\Delta_f^{A,B} < \Delta_f^{A,C}\ \&\ \Delta_g^{A,B} < \Delta_g^{A,C}$ then
          $success \leftarrow success + 1$
      end if
      if $\Delta_f^{A,B} > \Delta_f^{A,C}\ \&\ \Delta_g^{A,B} > \Delta_g^{A,C}$ then
          $success \leftarrow success + 1$
      end if
  end while
  $corr_{triplet} \leftarrow 2 \cdot \frac{success}{count} - 1$
```

While some correlations between our individual sub-models can be perceived (see Fig. 7), there does not seem to be a significant difference between correlations nor an interesting pattern. This is especially significant when comparing this plot with Fig. 4.

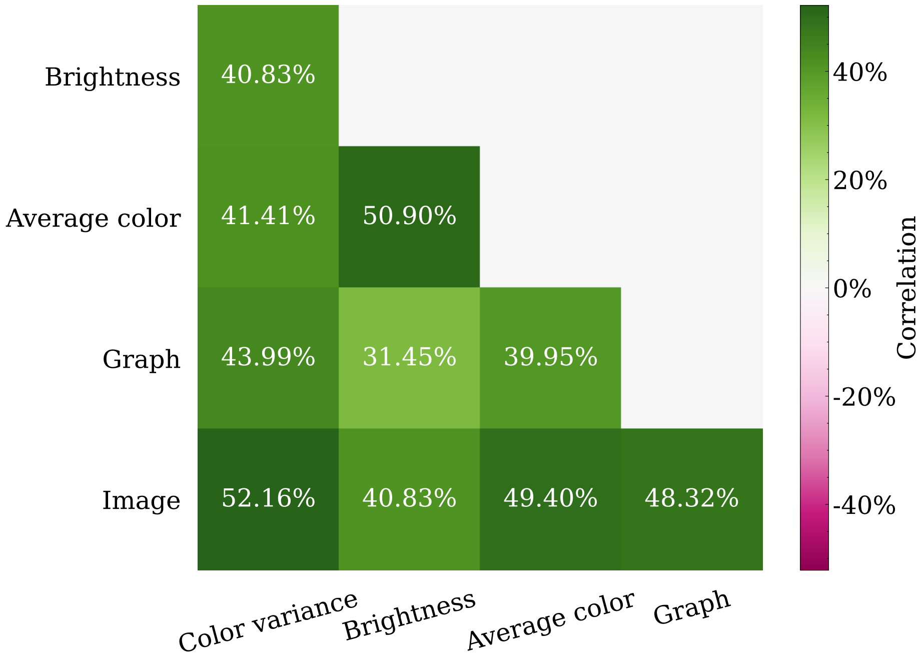

Fig. 7. Correlation (as defined in Alg. 1) between sub-model predictions.

3 Model specifications

In this subsection we include some hyperparameters that are interesting to reproduce our work, but where we did not find space for in the main paper

3.1 Image based submodel

Since the original images in the dataset are high-resolution images, they are resized to a dimension of 400 × 230 px. The resolution of the original images varies, since they have been automatically cropped based on YOLO's bounding boxes, as is described in [4]. Their average resolution, however, is 1814 × 1096 px.

The transformation in our approach consists of six convolutional layers, each increasing the number of features by 1.5 with a kernel size of 3. After every other convolution, we employ a max-pooling operator with a kernel size of 2 and add 3 dense layers to learn a final representation with 100 dimensions. We employ ReLU as an activation function, a learning rate of 0.001 and a batch size of 256.

4 Tabular version of figures in the main paper

This section provides further tables that correspond to figures included in the main body of work. They can provide interested readers with more details of the results and their standard deviations, and allow them to be further analyzed.

Table 1. Correlations (see Alg. 1) of sub-model predictions. The correlations are not quite symmetrical due to the use of random samples.

Model	Color variance	Brightness	Average color	Graph	Image
Color variance	-	40.5%	40.9%	43.5%	52.0%
Brightness	40.8%	-	51.3%	31.4%	41.1%
Average color	41.4%	50.9%	-	39.4%	48.6%
Graph	44.0%	31.4%	39.9%	-	47.5%
Image	52.2%	40.8%	49.4%	48.3%	-

Table 2. Tabular version of Fig. 3., showing the ranked accuracy of the ensembles and the best performing sub-model on [3].

Model	Rank-1 Accuracy	Rank-6 Accuracy	Rank-10 Accuracy
Concatenation	0.703 ± 0.079	0.92 ± 0.032	0.964 ± 0.02
Weighted Rank-1	0.675 ± 0.095	0.914 ± 0.033	0.952 ± 0.012
Weighted Triplet	0.49 ± 0.043	0.731 ± 0.046	0.811 ± 0.063
Majority Vote	0.37 ± 0.033	0.584 ± 0.02	0.627 ± 0.016
NN Triplet	0.331 ± 0.047	0.663 ± 0.053	0.759 ± 0.058

Table 3. Tabular version of Fig. 4, showing the improvement in Rank-1 accuracy of pairs of models over their expectation.

Model	Color variance	Brightness	Average color	Graph	Image
Color variance	-	5.7%	8.7%	6.2%	7.2%
Brightness	5.7%	-	12.6%	1.6%	2.4%
Average color	8.7%	12.6%	-	-1.4%	1.2%
Graph	6.2%	1.6%	-1.4%	-	9.1%
Image	7.2%	2.4%	1.2%	9.1%	-

Table 4. Tabular version of Fig. 10, showing the ranked accuracy of the ensembles and the best performing sub-model on [5].

Model	Rank-1 Accuracy	Rank-6 Accuracy	Rank-10 Accuracy
Weighted Rank-1	0.777 ± 0.054	0.975 ± 0.01	0.992 ± 0.007
Concatenation	0.698 ± 0.069	0.921 ± 0.032	0.958 ± 0.011
Majority Vote	0.598 ± 0.092	0.791 ± 0.06	0.813 ± 0.06
Weighted Triplet	0.349 ± 0.04	0.802 ± 0.043	0.902 ± 0.028
NN Triplet	0.349 ± 0.065	0.77 ± 0.036	0.866 ± 0.037

References

1. Ali, K., Pazzani, M.: Error reduction through learning multiple descriptions. Machine Learning **24** (11 1997)
2. Bian, Y., Chen, H.: When does diversity help generalization in classification ensembles? (10 2019)
3. Rutinowski, J., Chilla, T., Pionzewski, C., Reining, C., ten Hompel, M.: pallet-block-502 – A chipwood re-identification dataset (Sep 2021). https://doi.org/10.5281/zenodo.6353714
4. Rutinowski, J., Chilla, T., Pionzewski, C., Reining, C., ten Hompel, M.: Towards Re-Identification for Warehousing Entities – A Work-in-Progress Study. In: Proceedings of the IEEE Conference on Emerging Technologies In Factory Automation (ETFA). pp. 501–504 (2021)
5. Rutinowski, J., Endendyk, J., Reining, C., Roidl, M.: galvanized-636 – A galvanized steel re- identification dataset (Dec 2022). https://doi.org/10.5281/zenodo.7386956