MASKED SWINGERS: HARNESSING DATA AUGMENTATION TO ADVANCE AUTOENCODERS FOR SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) removes the need for annotations and makes models that are capable across more domains than supervised learning. The autoencoder SSL framework learns by reconstructing its own input after information loss through a bottleneck or noise injection. Masked autoencoders (MAE) are the most successful instantiation of this framework: they encode a random subset of patches, then decode the masked-out patches. In this work, we introduce key modifications to improve MAEs. Our method augments an image in two different ways, then masks and encodes each view separately. It then *exchanges the global representations (`CLS` tokens) between views* before decoding the masked patches. By design, our **Masked Swingers** encourages learning a view-agnostic summary of the image to facilitate efficient transfer. We perform extensive experiments, and find Masked Swingers outperforms MAE by +3–5% on ImageNet-1K kNN and provides large gains on fine-grained tasks, e.g., relative gains of +45% on instance retrieval, +22% on animal re-ID, and +76% on Omniglot character recognition. To boot, Swingers reduces error −64% relative to MAE on three new state-probing datasets, opening the door to world modeling. Welcome to our Swingers party.

1 INTRODUCTION: AUTOENCODING AND VIEWS

Autoencoding (Cottrell et al., 1987; Ballard, 1987) is an old, yet still popular self-supervised learning (SSL) framework. It has an encoder, which receives an input and hands a hidden state to a decoder, which predicts the input—and to learn *useful* representations, it puts an obstacle somewhere along the way, e.g., by masking encoder-input portions (Vincent et al., 2008; Pathak et al., 2016; He et al., 2022) or sending the hidden state through a low-dimensional bottleneck to learn a summary representation (Bourlard & Kamp, 1988; Hinton & Salakhutdinov, 2006; Van Den Oord et al., 2017).

Joint embedding (Becker & Hinton, 1992; Bromley et al., 1993) is another popular SSL framework. It separately encodes two views of a sample and encourages their representations to be similar—and to learn *useful* representations, it controls the mutual information between the views, often through data augmentation (Chen et al., 2020; He et al., 2020). Because the loss is computed on a single vector per view, the encoder learns to summarize its input. This insight, that two views of a sample should be represented similarly, underpins many SSL algorithms (Chen et al., 2021; Bardes et al., 2021), yet no autoencoding method directly encourages learning view-agnostic representations.

We thus introduce **Masked Swingers**, which harnesses data augmentation to improve the representations learned from autoencoding. Our Masked Swingers algorithm augments and masks a sample in two different ways and encodes each view. It then exchanges the encoded "`CLS`" tokens between views—exchanging messages between them—and decodes the originally masked portions of each view. This encourages the encoder to summarize its input to help reconstruct an arbitrary view through a bottleneck (Tishby et al., 2000). Because our method augments and masks views independently, it encourages this summary to be useful regardless of the view-specific transformations.

Masked Autoencoders (MAE; He et al., 2022) offer several advantages. They have proven effective across data distributions and modalities, including natural imagery (Singh et al., 2023), satellite imagery (Cong et al., 2022), 1D time series (Li et al., 2023), and 3D point clouds (Pang et al., 2022). They have low device-memory requirements: only one sample at a time must fit in memory for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

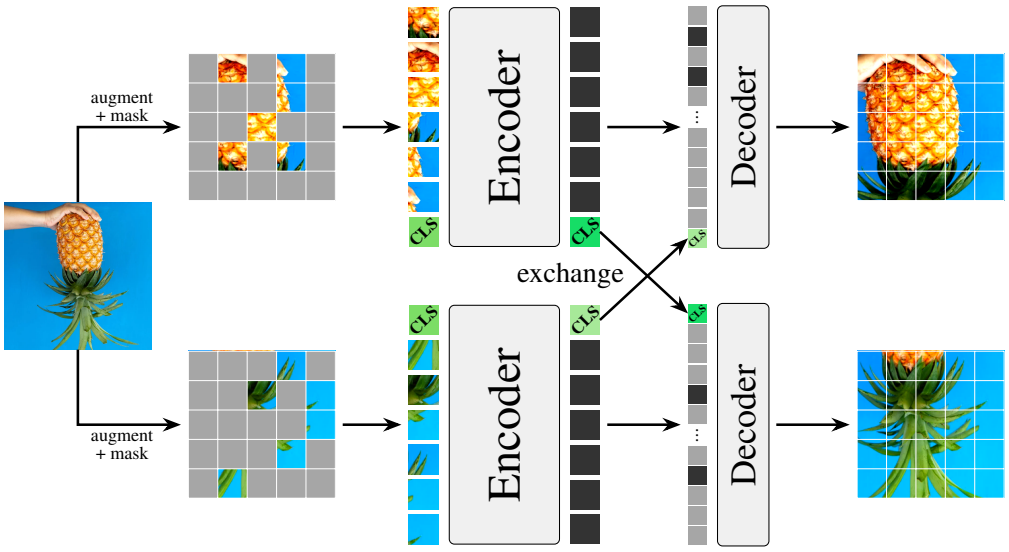


Figure 1: Our **Masked Swingers** exchanges **CLS tokens** between different views of the same image prior to reconstruction. We make different views through independent data augmentation and masking. Together, these two design choices encourages the encoder to summarize its input to help reconstruct an arbitrary view so that it can be transferred efficiently and broadly.

training. This is because, unlike most joint-embedding methods, autoencoders need no cross-sample negatives or statistics to avoid collapse, and their Vision Transformer backbone (ViT; Dosovitskiy et al., 2021) does not use batch normalization (Ioffe & Szegedy, 2015). MAE pre-training also tends to be robust because its targets are stationary/grounded, unlike self-distillation methods (Assran et al., 2023). MAE is thus frequently used to pre-train ViTs across settings, so we aim to improve it.

We now summarize our contributions to start the party. **(i)** We propose a novel autoencoding SSL algorithm called Masked Swingers, which harnesses data augmentation to encourage a view-agnostic image summary for efficient and general transfer. **(ii)** We extensively show our Masked Swingers outperforms MAE across 67 datasets when used as a frozen feature extractor—gains are especially large on fine-grained tasks, e.g., instance retrieval (+45% relative gain), animal re-identification (+22% relative gain), and species classification (+43% relative gain). **(iii)** We study mask rate and model size, showing that Swingers can benefit from lower mask rates than MAE and that our method scales well across Tiny/Small/Base/Large ViTs. **(iv)** We find that we can exchange other tokens, besides the CLS, to still perform well. **(v)** We show that Swingers increases the effective rank of encoders, using the embedding space more evenly. **(vi)** We provide preliminary results in world modeling—Swingers excels at state prediction, reducing error rate −64% relative to MAE.

2 Background: Masked Autoencoders

We review details of the MAE algorithm because our Masked Swingers inherits much of its design.

Patchifying. An image $\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$ is divided into $N=HW/P^2$ non-overlapping patches of resolution $P\times P$ and flattened into a sequence $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N]$, $\boldsymbol{x}_i\in\mathbb{R}^{P^2C}$. A linear layer maps these patches to the token dimension d, which is the encoder "width", and position embeddings $\boldsymbol{p}_i\in\mathbb{R}^d$ are added to make patch tokens $\boldsymbol{t}_i=\mathrm{Linear}(\boldsymbol{x}_i)+\boldsymbol{p}_i$.

Masking. A random subset $\mathbb{M}$ of ρN patch indices is dropped, leaving the *visible* patches $\mathbb{V}$, where ρ is the masking ratio. MAE samples masks uniformly; ColorMAE (Hinojosa et al., 2024) thresholds uniform noise passed through a 2D filter, giving more contiguous visible and masked regions. "Red" color masking yields large contiguous masks and "green" masking yields smaller regions. Because of spatial autocorrelation, for the same mask rate, red and green color masking result in less mutual information between independently sampled masks of the same image, relative to uniform masking.

Encoding. A learnable CLS token $\boldsymbol{t}_{\texttt{CLS}}$ is prepended to the visible tokens. Crucially, the encoder f_θ is applied only to these tokens, $\left[\boldsymbol{z}_{\texttt{CLS}}, \{\boldsymbol{z}_i\}_{i\in\mathbb{V}}\right] = f_\theta\left(\left[\boldsymbol{t}_{\texttt{CLS}}, \{\boldsymbol{t}_i\}_{i\in\mathbb{V}}\right]\right), \boldsymbol{z}_i \in \mathbb{R}^d$.

Decoding. The decoder operates on all N patch positions plus the CLS token. Its inputs $\boldsymbol{h}_i \in \mathbb{R}^{d'}$ are $\boldsymbol{h}_i = \text{Linear}(\boldsymbol{z}_i) + \boldsymbol{p}'_i$ for visible patches $i \in \mathbb{V}$, projecting to the (narrower) decoder width d', and $\boldsymbol{h}_i = \boldsymbol{m} + \boldsymbol{p}'_i$ for masked patches $i \in \mathbb{M}$, where $\boldsymbol{m} \in \mathbb{R}^{d'}$ is a learnable mask token that is shared across *all* masked positions and $\boldsymbol{p}'_i$ are decoder position embeddings. The CLS token uses the same linear layer, $\boldsymbol{h}_{\texttt{CLS}} = \text{Linear}(\boldsymbol{z}_{\texttt{CLS}})$. Transformer blocks, g_ϕ, map these to latents, $[\boldsymbol{u}_{\texttt{CLS}}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_N] = g_\phi\left([\boldsymbol{h}_{\texttt{CLS}}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]\right)$. We discard $\boldsymbol{u}_{\texttt{CLS}}$ and linearly map each patch latent to pixel predictions, $\hat{\boldsymbol{x}}_i = \text{Linear}(\boldsymbol{u}_i) \in \mathbb{R}^{P^2C}$.

Objective. The loss, $\mathcal{L}_{\text{MAE}}(\theta, \phi) = \frac{1}{|\mathbb{M}|}\sum_{i\in\mathbb{M}} \|\hat{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_i\|_2^2$, is the mean squared error on masked positions only. Targets are pixels, $\tilde{\boldsymbol{x}}_i = (\boldsymbol{x}_i - \mu_i)/\sigma_i$, normalized by the mean μ_i and std σ_i of patch i.

3 Masked Swingers: Encouraging a useful CLS token

Masked Swingers changes one thing about the MAE in Sec. 2: the CLS token the decoder receives is from the encoder of a *different* view, created via separate masking and data augmentation (Fig. 1).

Two views. We augment each image twice, into $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, which we patchify and mask independently, giving visible sets $\mathbb{V}^1, \mathbb{V}^2$ and masked sets $\mathbb{M}^1, \mathbb{M}^2$. We encode the two views separately by the same encoder f_θ, producing $\boldsymbol{z}^1_{\texttt{CLS}}, \boldsymbol{z}^2_{\texttt{CLS}}$ and patch tokens $\{\boldsymbol{z}^1_i\}_{i\in\mathbb{V}^1}, \{\boldsymbol{z}^2_i\}_{i\in\mathbb{V}^2}$.

Swinging. We decode the input exactly as in MAE, from each view's own patch tokens and mask tokens, with one key change: we trade the CLS tokens between views, $\boldsymbol{h}^1_{\texttt{CLS}} = \text{Linear}(\boldsymbol{z}^2_{\texttt{CLS}})$ and $\boldsymbol{h}^2_{\texttt{CLS}} = \text{Linear}(\boldsymbol{z}^1_{\texttt{CLS}})$. Everything downstream is unchanged from MAE: we map each sequence to latents with g_ϕ, we discard the CLS, and we project the remaining tokens back to pixels.

Objective. We compute the loss as the MAE reconstruction error of Sec. 2, averaged over both views, $\mathcal{L}_{\text{Swingers}}(\theta, \phi) = \frac{1}{2}\sum_{v\in\{1,2\}} \frac{1}{|\mathbb{M}^v|}\sum_{i\in\mathbb{M}^v} \|\hat{\boldsymbol{x}}^v_i - \tilde{\boldsymbol{x}}^v_i\|_2^2$.

Since the masked patches of $\boldsymbol{x}^1$ must be inferred from its own visible patches and the CLS token of $\boldsymbol{x}^2$ (and symmetrically), the encoder must compress each view into a summary that stays useful under a different augmentation and a different mask—promoting a view-agnostic summary. We call this algorithm "**Full-Swingers**", to distinguish it from the Masked-Swinger variant introduced next.

Half-Swingers. We propose a simple variant that uses the CLS token from one view to decode both views (equivalent to exchanging CLS tokens on 50% of images and keeping the same-view CLS tokens for the other 50%), $\boldsymbol{h}^1_{\texttt{CLS}} = \text{Linear}(\boldsymbol{z}^1_{\texttt{CLS}})$ and $\boldsymbol{h}^2_{\texttt{CLS}} = \text{Linear}(\boldsymbol{z}^1_{\texttt{CLS}})$. Our "Half-Swingers" trains the encoder to simultaneously **(i)** summarize its own view to help decode arbitrary views and **(ii)** infer the masked patches of its own view. The encoder is not aware if its CLS helps reconstruct its own view or another view, and the decoder is not aware from which view its CLS originates—so the encoder may compromise between "compressing the visible" and "inferring the missing".

4 Experimenting Widely with Controls and Context

In Sec. 4.1, we run controlled experiments to compare our Masked Swingers to MAE, avoiding confounds that could otherwise account for performance gaps. We report the performance of models pre-trained using other frameworks for reference—not meant for direct comparison. Since self-supervised representations should be general, we evaluate on ImageNet-1K (Russakovsky et al., 2015), the gold-standard, alongside several other datasets that check model performance for other uses, e.g. instance retrieval (Kordopatis-Zilos et al., 2025; Radenović et al., 2018; Oh Song et al., 2016) and fine-grained classification (Van Horn et al., 2021; Picek et al., 2025). We also fine-tune on ImageNet-1K to check if our Masked Swingers retains the strong fine-tunability of MAE.

In Sec. 4.2, we study the effects of masking rate and model size on both Swinger variants because these are two important design axes. Then in Sec. 4.3, we provide: **(i)** analysis, by looking at data augmentation's interaction with pre-training length, and the relationship between embedding effective rank (or usage) and accuracy; **(ii)** ablations, by giving both encoders the same augmentation to confirm our motivating hypothesis; and **(iii)** alternatives, by exchanging tokens besides the CLS.

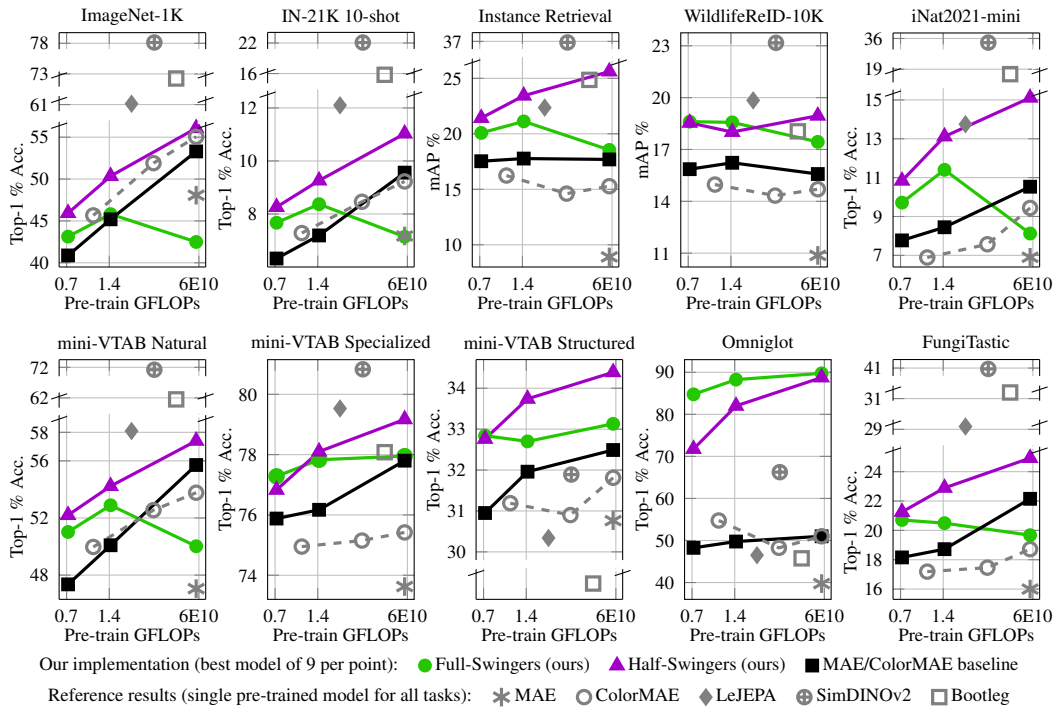


Figure 2: **Our Half-Swingers extracts better features than MAE** evaluated over 10 tasks and 3 pre-training lengths. Our **Full-Swingers** beats MAE on natural image classification (IN-1K, IN-21K, and mini-VTAB natural) on *shorter* schedules, and all other tasks on most schedules. In our controlled experiments (●▲■), we run 9 pre-trainings (varying data augmentation and color masking strengths) per SSL algorithm and pre-training length, choosing the best config per point. Intended for reference only, we also run evaluations on public/official pre-trained checkpoints (✱○♦⊕□).

Finally in Sec. 4.4, we explore if Masked Swingers can help with world modeling, as a new and exciting use of pre-trained vision encoders. We do this by probing the features extracted by models for "state" information on three new datasets (Karmann et al., 2026).

4.1 Pre-training: Varying Augmentation, Color Masking, and Epochs

Pre-training Setup: Varied Settings. We compare three autoencoding SSL algorithms: the MAE/ColorMAE baseline and our two Masked-Swinger variants, Full-Swingers and Half-Swingers. For each algorithm, we pre-train 27 models spanning the cross product of pre-training length {100, 200, 800 epochs}, data augmentation strength {weak, medium, strong}, and color masking {none, green, red}. Our weak, medium, and strong settings apply RandAug (Cubuk et al., 2020) with progressively more operations at higher magnitudes. We vary augmentation and color masking because they control the shared information between views, which is important for Swingers. We also vary these knobs for the MAE baseline to equally tune it, ensuring experimental fairness.

Pre-training Setup: Fixed Settings. We use the ViT-B/16 architecture (Dosovitskiy et al., 2021), which is the most common ViT size, and study other model sizes in Sec. 4.2. We pre-train on ImageNet-1K at 224×224 resolution. Since Swingers encodes two views per image, we train the MAE baseline with a repeat augmentation (Berman et al., 2019; Hoffer et al., 2020) of 2, i.e., each image appears twice per batch under independent augmentations and masks. All three algorithms thus process the same number of views per step, so we hold the batch size, learning rate, and total pre-training compute fixed across them; the algorithm implementations differ by only a few lines of code. For all other settings, including optimization hyperparameters, we use MAE's defaults, since Masked Swingers shares MAE's reconstruction objective and should have similar optimal values.

Other models as references. To contextualize our controlled experiments, we also evaluate publicly released models, which differ from ours in their pre-training recipes and are not meant for direct

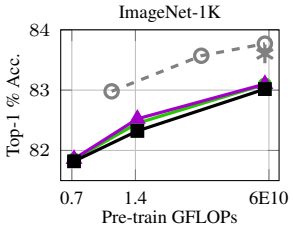


Figure 3: **Our Masked Swingers are as fine-tunable as the MAE baselines.** In our controlled setting, Half-Swingers (▲) edges Full-Swingers (●), which edges the MAE/ColorMAE baseline (■). However, pre-training length affects downstream accuracy much more than these differences. The reported results from He et al. (2022) (✳) and Hinojosa et al. (2024) (○) outperform these runs, yet they differ with ours in their number of pre-training steps per FLOP and implementations.

comparison. First, we evaluate the official MAE and ColorMAE checkpoints. Since they were pre-trained without repeat augmentation, they take twice as many optimization steps per FLOP as our controlled MAE baselines. Second, we evaluate three strong recent *non-autoencoding* models: LeJEPA (Balestriero & LeCun, 2025), SimDINOv2 (Wu et al., 2025), and Bootleg (Lowe et al., 2026). Like ours, these models use ViT-B/16 and pre-train on ImageNet-1K at 224×224 resolution. LeJEPA has no public ViT-B/16 checkpoint, so we use a public reproduction (Appendix A.3).

Evaluation Setup: Frozen Features. We evaluate frozen features on 10 tasks spanning 67 datasets: **(i)** kNN on ImageNet-1K (Russakovsky et al., 2015); **(ii)** kNN on ImageNet-21K (Ridnik et al., 2021) with limited labels, using 10 training images from each of its 10,450 classes (104.5K images) and all 522.5K validation images; **(iii)** kNN on mini-VTAB (Fuller et al., 2026), which is a more feasible version of VTAB (Zhai et al., 2019) with 21 tasks in total, split into its **(a)** natural, **(b)** specialized, and **(c)** structured subsets; **(iv)** instance retrieval on Revisited Oxford and Paris (Radenović et al., 2018) (ROxford5k and RParis6k, medium protocol), Stanford Online Products (Oh Song et al., 2016), and the ILIAS core set (Kordopatis-Zilos et al., 2025), without YFCC100M distractors (Thomee et al., 2016); **(v)** animal re-identification on WildlifeReID-10k (Adam et al., 2025), which covers 10.7K individuals from 37 datasets, using the standard open-set split with a single pooled gallery; **(vi)** handwritten character recognition on Omniglot's 20-way one-shot benchmark (Lake et al., 2015); **(vii)** kNN on iNaturalist 2021-mini (Van Horn et al., 2021), with 10K species, 50 training images each (500K total), and all 100K validation images; and **(viii)** kNN on the long-tailed closed-set split of FungiTastic (Picek et al., 2025), with 2.8K fungal species, 433.7K training images, and 89.7K validation images. We report mAP for instance retrieval and animal re-identification, and top-1 accuracy for all other tasks. For kNN, we tune the number of neighbors and the softmax temperature per model and dataset. We evaluate both the `CLS` token and the mean-pooled patch tokens, and report the better of the two on each dataset (see Appendix Sec. A.4). Finally, for each algorithm and pre-training length, we report the best of the 9 augmentation and color-masking settings on each task; this selection is applied identically to MAE and Swingers so the comparisons are fair.

Results: Masked Swingers learns better frozen features than MAE on all 10 tasks (Fig. 2). On IN-1K kNN, Half-Swingers outperforms MAE by +5.1/+5.2/+2.8 percentage points (pp) at 100/200/800 pre-training epochs. Gains are larger on the harder 10-shot IN-21K task, which has more classes and fewer examples per class: Half-Swingers improves accuracy by +30.7%/+28.8%/+15.4% relative to MAE. Our gains tend to be larger still on the remaining 8 tasks. On Omniglot at 800 epochs, for example, both Swinger variants reach around 90% accuracy versus 51% for MAE, approaching the ∼95% of specialized few-shot methods (Lake et al., 2015). These results select the best of 9 settings per point, but Half-Swingers also outperforms MAE when averaged over all 9 augmentation and color-masking settings, by +5.6/+8.0/+0.5 pp on IN-1K kNN. The trend with pre-training length depends on the task. On natural-image tasks (IN-1K, IN-21K, and mini-VTAB natural), our gains are largest for shorter pre-training, whereas on the remaining tasks they maintain or grow as pre-training lengthens. Please see Appendix A.5 for all results.

Evaluation Setup: Fine-tuning. We fine-tune all 81 pre-trained models (3 algorithms $\times$ 3 pre-training lengths $\times$ 9 augmentation and color-masking settings) on IN-1K for 100 epochs, tuning the learning rate for each. As with frozen features, we report top-1 accuracy for the best of the 9 settings per algorithm and pre-training length. Please see Appendix A.2 for more fine-tuning details.

Results: Masked Swingers retains MAE's fine-tunability (Fig. 3). In our controlled experiments, both Swinger variants fine-tune at least as well as MAE on IN-1K. None of our fine-tuned models, including our MAE baselines, reach the IN-1K accuracies reported by He et al. (2022) and Hinojosa et al. (2024). Fine-tuning implementations differ between these references and ours, e.g., we use

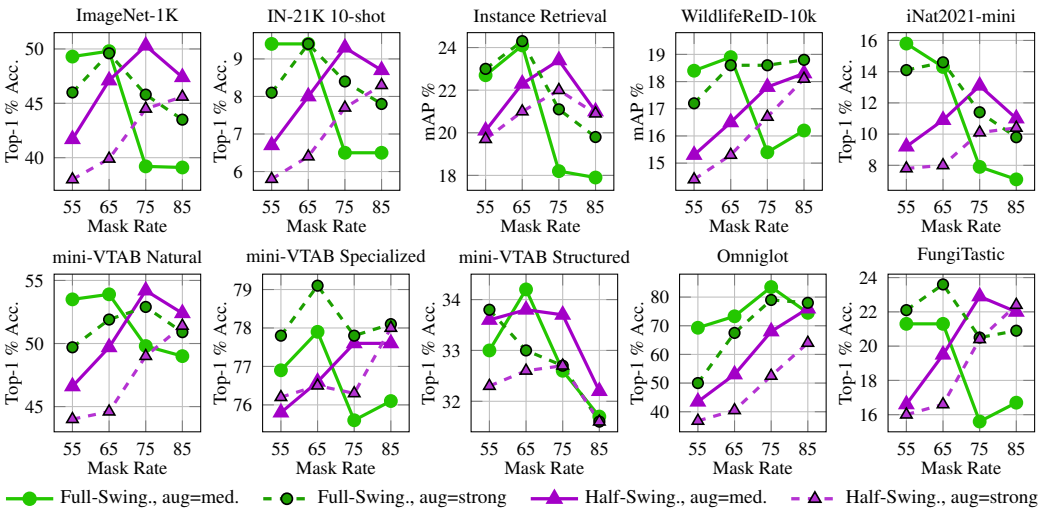


Figure 4: **Full-Swingers prefers lower mask rates than Half-Swingers.** For example, at a 65% mask rate, our Full-Swingers ties or outperforms Half-Swingers at its best mask rate of 75%. Omniglot is an exception where higher mask rates improves Full-Swingers. Stronger augmentation tends to make Full-Swingers more robust to mask rates and it tends to hurt Half-Swingers.

JAX (Bradbury et al., 2018) and TPUs (Jouppi et al., 2023). Pre-training also differs. In particular, Fig. 3 compares models at equal pre-training FLOPs (x-axis), and at equal FLOPs the reference models take twice as many optimization steps as ours, since they do not use repeat augmentation.

4.2 Pre-training: Varying Mask Rate and Model Size, alongside Augmentation

Pre-training Setup and Evaluations. We pre-train Full and Half-Swingers for 200 epochs on ImageNet-1K without color masking. We run two sweeps, each crossing augmentation strength {medium, strong} with a second design axis to study their interaction: masking ratio {55%, 65%, 75%, 85%} using the ViT-Base size, and model size {Tiny, Small, Base, Large} using a 75% masking ratio. We use the same frozen-feature evaluations as in Sec. 4.1.

Results: Lower mask rates help Full-Swingers (Fig. 4). Full-Swingers performs better with lower mask rates than MAE's default of 75%. At a 65% mask rate, it reaches 50% IN-1K kNN accuracy, matching Half-Swingers' best result. Compared with the default 75%, this lower mask rate also improves Full-Swingers on 10-shot IN-21K, all three mini-VTAB subsets, instance retrieval, animal re-identification, and fine-grained species classification. The exception is Omniglot, where accuracy drops from 83% to 73%, which may be a worthwhile trade-off given the gains elsewhere. Stronger augmentation also makes Full-Swingers more robust to the choice of mask rate. In contrast, Half-Swingers performs best or near-best at the default 75% mask rate. Omniglot is again the exception: raising the mask rate to 85% improves accuracy by +8/+12 pp with medium/strong augmentation.

Results: Full-Swingers wins at smaller ViT sizes and Half-Swingers at larger sizes (Fig. 5). Both variants generally improve with model size. Full-Swingers outperforms Half-Swingers at smaller sizes, but Half-Swingers scales better to ViT-L. This comparison uses the default 75% mask rate, however, which is suboptimal for Full-Swingers (Fig. 4), so a lower mask rate may narrow the gap at larger sizes. Stronger augmentation tends to help Full-Swingers across sizes. For Half-Swingers, the pattern is more mixed: stronger augmentation helps for ViT-Tiny and ViT-Large.

4.3 Analysis, Ablations, and Alternatives

Full-Swingers can lose kNN accuracy on ImageNet with longer pre-training. In Fig. 2, Full-Swingers' kNN accuracy on IN-1K and IN-21K increases from 100 to 200 epochs but decreases from 200 to 800 epochs. Because each point is the best of 9 settings, this curve does not track any single setting, and the decrease does not occur in every setting. To look closer, Tab. 1 shows how

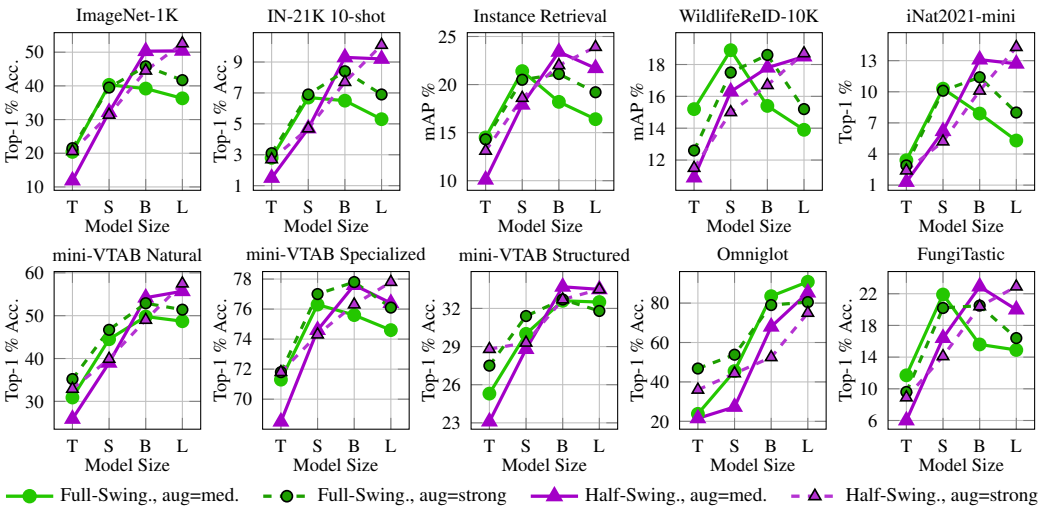


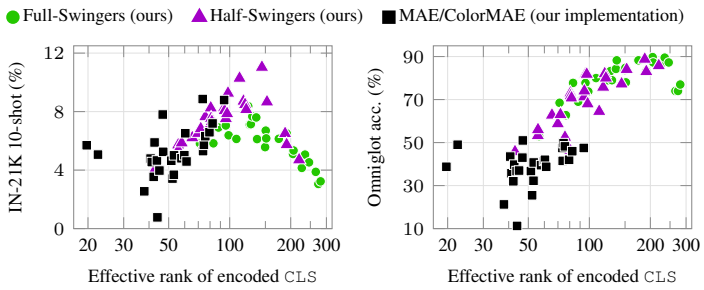
Figure 5: **Full-Swingers and Half-Swingers tend to improve when increasing model size.** The gain is monotonic for Omniglot but is more complex for other tasks. For example, Full-Swingers (aug=medium) peaks on animal re-identification at ViT-Small, then decreases for ViT-Base and ViT-Large; increasing augmentation helps for these larger architectures.

Table 1: **Full-Swingers need stronger augmentation when pre-trained for longer.** Results are ImageNet-1K kNN top-1 % accuracy. "Red" masks larger contiguous areas, "None" is uniform sampling as in MAE.

Aug. Strength	Color Masking	Pre-training epochs 100	200	800
Weak	None	39.1	33.9	33.2
Strong	Red	30.4	37.1	41.2

lengthening pre-training affects two settings at opposite ends of our design space: **(i)** weak augmentation without color masking, which produces the most similar views, and **(ii)** strong augmentation with red color masking, which produces the least similar views. With longer pre-training, IN-1K kNN accuracy decreases when the views are too similar and increases when they are more dissimilar. The mutual information between views thus shapes the representations our method learns.

Figure 6: **Swingers distributes the variation of the `CLS` token more evenly across directions.** Effective rank of the `CLS` covariance on IN-1K val vs. (left) 10-shot IN-21K kNN accuracy and (right) Omniglot 20-way one-shot acc., for 81 pre-trained models (3 aug strengths × 3 masking strengths × 3 training lengths × 3 algorithms).



Swingers' `CLS` token varies along more directions. In MAE, the decoder predicts masked patches from the visible patch tokens of the same view, so the `CLS` token is not directly encouraged to carry information. In Swingers, the `CLS` token is the only information about the other view that the decoder receives, so the more descriptive it is, the lower the reconstruction loss. To measure this, we compute the effective rank (Roy & Vetterli, 2007) of the `CLS` covariance on the IN-1K validation set, which measures how evenly the variation across images is spread over the embedding's directions: it equals the embedding dimension when every direction contributes equally, and approaches 1 when a single direction dominates. Fig. 6 plots effective rank against 10-shot IN-21K and Omniglot accuracy for all 81 models from Sec. 4.1. On IN-21K, accuracy rises with rank up to a point and then declines, whereas on Omniglot it rises steadily with rank. One possible explanation is that the

additional directions capture "lower-level" features, such as the subtle stroke shapes that distinguish Omniglot's 1,623 handwritten characters, which help Omniglot but eventually hurt IN-21K.

Table 2: **Swingers needs independently augmented views to learn "higher-level" frozen features.** Encouraging view-*specific* representations, by sharing augmentations between views, drops IN-1K kNN by −28% and −38%, for Full and Half-Swingers, respectively. Yet it can boost downstream fine-tuning accuracy. "CLS" and "Pool" results are computed via kNN. **Best** and $\underline{2^{nd}\text{ best}}$ per column.

Same Aug.	Swing Strategy	ImageNet-1K CLS	ImageNet-1K Pool	ImageNet-1K FT	Omniglot CLS	Omniglot Pool
Yes	None	38.0	<u>37.9</u>	81.4	47.3	<u>40.8</u>
	Half	12.1	29.6	<u>81.8</u>	38.5	36.0
	Full	11.7	24.1	**82.0**	43.5	33.8
No	None	35.3	**38.1**	81.3	41.5	38.3
	Half	**50.3**	37.8	81.5	<u>68.0</u>	37.5
	Full	<u>39.2</u>	35.6	81.5	**83.3**	**47.3**

Swingers needs independent augmentation for strong frozen features (Tab. 2). We ablate independent augmentation by giving both views the same augmentation, so they differ only in their masks. This substantially hurts both variants, reducing IN-1K kNN accuracy by 25.2 pp for Full-Swingers and 38.0 pp for Half-Swingers. Shared augmentation also hurts Omniglot accuracy. These results show that independent augmentation encourages more view-agnostic or "semantic" representations. Full-Swingers *without* independent augmentation achieves our best IN-1K fine-tuning accuracy (82.0%). In this setting, the exchanged CLS token describes a view that differs only in its mask, which encourages a view-*specific* summary that may suit fine-tuning slightly better.

Table 3: **Swinging other tokens can also work well, e.g., exchanging both CLS and pooled patches.** Exchanging patch-pooled tokens may not need to be half-swung under this setting to perform well (i.e., 47% ImageNet-1k kNN). **Best** and $\underline{2^{nd}\text{ best}}$ per column.

Setting	IN-1K CLS	IN-1K Pool	Omniglot CLS	Omniglot Pool
exchange CLS (Full-Swingers)	39.2	35.6	83.5	46.8
exchange CLS on 50% of imgs (Half-Swingers)	<u>50.4</u>	37.8	68.0	37.3
exchange pooled patches	42.8	**46.8**	60.0	**80.5**
exchange pooled patches on 50% of imgs	38.4	<u>44.8</u>	51.5	55.0
exchange CLS and pooled patches	36.3	36.8	80.0	<u>64.0</u>
exchange CLS and pooled patches on 50% of imgs	**50.5**	39.1	59.5	47.5
provide both CLS to both decoder views	40.3	35.1	<u>83.8</u>	44.8
↪ add two position embeddings	43.0	36.1	**86.8**	42.5

We can also exchange pooled patch tokens (Tab. 3). We test exchanging tokens other than the CLS token. In one alternative, each decoder receives its own view's CLS token and patch tokens and the mean-pooled patch token from the other view. This is competitive with exchanging CLS tokens, showing that the benefit comes from passing a cross-view summary rather than from the CLS token specifically. We also exchange both the pooled patch token and the CLS token. This combination works best when exchanging occurs on only 50% of samples, as in Half-Swingers.

4.4 State Probing for World Modeling

Table 4: **Masked Swingers shows promise for world modeling.** Masked Swingers improves state probing NMSE (lower ↓ is better) on the three MotionJEPA games (Karmann et al., 2026). And this improvement is *large*, e.g., 64% error reduction from MAE/ColorMAE to our Full-Swingers (averaged over all tasks and models). The MotionJEPA result is *not* directly comparable: its encoder is ViT-T/14 at 112^2px trained from scratch on each game, whereas all other models are ViT-B/16 at 224^2px pre-trained on ImageNet-1K. **Best** and $\underline{2^{nd}\text{ best}}$ per column.

Model	Pong	Dino	Golf
Pre-trained by us on ImageNet-1K			
MAE/ColorMAE (mean)	0.044	0.058	0.091
MAE/ColorMAE (best)	0.007	0.031	<u>0.033</u>
Half-Swingers (mean)	0.012	0.032	0.054
Half-Swingers (best)	<u>0.005</u>	<u>0.013</u>	**0.024**
Full-Swingers (mean)	0.009	0.023	0.036
Full-Swingers (best)	<u>0.005</u>	**0.012**	**0.024**
Pre-trained by others on ImageNet-1K			
SimDINOv2	0.006	0.044	0.059
Bootleg	0.008	0.085	0.061
LeJEPA	0.014	0.052	0.102
ColorMAE (best of 3)	0.011	0.033	0.038
MAE	0.034	0.035	0.123
Trained by Karmann et al. (2026) on the games themselves			
MotionJEPA	**0.004**	0.021	0.061

Masked Swingers reduces state-prediction error by 64% relative to MAE. We probe frozen features for state information on three tasks, Pong, Dino, and Golf, using the public codebase of

Karmann et al. (2026). We evaluate our models from Sec. 4.1 along with the reference models. Our Swingers substantially outperforms the MAE/ColorMAE (Tab. 4), reducing error by 64% averaged over all models and tasks. It also outperforms SimDINOv2, reducing error by 73% on Dino and 59% on Golf. These tasks require low-level features, such as the character's height or an obstacle's position. Swingers' strength here is therefore consistent with its gains on other low-level tasks in Sec. 4.1, such as the mini-VTAB structured subset and Omniglot handwritten character recognition.

5 Related Work: The Parties of Friends and Family

Cross-View Autoencoding. The closest method to ours is CropMAE (Eymaël et al., 2024), designed for video feature-correspondence tasks such as DAVIS (Pont-Tuset et al., 2018). CropMAE makes a second view with a mild crop and a random horizontal flip, then masks nearly all of it (ρ=98.5%). It encodes the unmasked original view and the masked second view separately, then decodes the masked patches using all tokens from both views. CropMAE paper does not report IN-1K, so we evaluate its pre-trained ViT-S checkpoint ourselves: it reaches 12% IN-1K kNN accuracy, compared with 40% for our Swingers ViT-S (Fig. 5). CroCo (Weinzaepfel et al., 2022) and SiamMAE (Gupta et al., 2023) are the next closest methods, but they obtain their views from data rather than augmentation: CroCo uses different views of a 3D scene and SiamMAE uses neighboring video frames. CroCo and CroCo v2 (Weinzaepfel et al., 2023) reach 19% and 29% ViT-B IN-1K kNN accuracy, though they pre-train on different data for different lengths, so these numbers are not directly comparable to ours. Unlike these methods, which let the decoder access every token of the other view, we pass *one* token between views. This information bottleneck, combined with strong, independent augmentation and masking of each view, encourages a view-agnostic input summary.

Other Masked Autoencoders. Many papers, including ColorMAE (Hinojosa et al., 2024), invent clever masking strategies (Li et al., 2022a; Kakogeorgiou et al., 2022; Shi et al., 2022; Shin et al., 2024; Choi et al., 2024; Wang et al., 2023a; Feng & Zhang, 2023; Yang et al., 2026; Li et al., 2022b; Liu et al., 2023b; Haghighat et al., 2024). Others create new image-derived targets (Wei et al., 2022; Xie et al., 2023; Wang et al., 2023b; Liu et al., 2023a). And some papers improve the decoder (Fu et al., 2025; Rajasegaran et al., 2025; Wang et al., 2023b). These may provide complementary improvements to our Swingers. Since masking affects the mutual information between views and could thus join our Swingers party, we thoroughly experiment with ColorMAE's masking (Sec. 4.1). We choose ColorMAE over alternatives because it is data-independent, model-free, and intuitive.

Masked Latent Modeling. There is a related but distinct SSL framework which masks inputs, then predicts representations (not pixels, thus are not autoencoders) at the masked locations (Assran et al., 2023; Lowe et al., 2026; Wei et al., 2024; Baevski et al., 2022; Zhou et al., 2021). The latents are computed from a "teacher" model that is the EMA of the "student", thus relies on self-distillation. These methods may join future and deeper Swingers parties to encourage view-agnostic summaries.

6 Swinging Shut

Limitations and future work. The main limitation of Masked Swingers is that it requires data augmentation to make different views, which is not required by MAE. However, data augmentation is ubiquitous across deep-learning applications, so this limitation is minor in practice. Swingers can underperform SOTA methods from other SSL frameworks, e.g., SimDINOv2. However, we narrow this gap while keeping the strengths of the autoencoding framework, i.e., robustness to training settings (by predicting stationary/grounded targets, e.g., images) and low device-memory requirements (by avoiding cross-sample interactions). And we achieve this with a simple and intuitive idea. The main limitation of our study is that we only pre-train on natural images. We choose this for focus, leaving other experiments to future Swingers parties in space (3D data) and time (video).

Conclusion. We introduce a novel autoencoding algorithm that encourages view-agnostic summary representations by exchanging encoded `CLS` tokens between two different views of an image before decoding. Via thorough controlled experiments, we show our Masked Swingers significantly outperforms the MAE baseline, especially on fine-grained tasks. Through ablations we show that independent augmentations during pre-training is crucial to extracting "higher-level" features. And we show our Swingers' potential as a frame-encoder in world modeling. The lights are off and the masks are on, but our Swingers parties have just begun—join us.

AI USE STATEMENT

In this work, we used generative AI tools for: **(i)** drafting the initial methods section, which we thoroughly edited; **(ii)** writing the entire appendix given the codebase, which we reviewed; **(iii)** suggesting different words in the main paper to improve clarity and brevity; **(iv)** suggesting related work that we may have missed; **(v)** creating pre-training configuration files, which we reviewed and logged; and **(vi)** coding the frozen-feature evaluations, given our previous codebase with similar functionality for feature extraction and kNN (originally written by us), which we reviewed. We estimate that $\sim$5% of the words in the main text were written by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our method is simple to implement starting from an MAE codebase. Since we use JAX and TPUs, we start with this ***public*** codebase: `https://github.com/KAIST-AILab/MaskedAutoencoder-Jax`. Upon publication, we will share all of our 81 pre-trained models, pre-training code, and frozen-feature evaluation code to support reproducibility and make it easier for others to join the party.

REFERENCES

Lukas Adam, Vojtech Cermak, Kostas Papafitsoros, and Lukas Picek. WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2124–2134, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pp. 1298–1312. PMLR, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. doi:10.48550/arxiv.2511.08544.

Dana H Ballard. Modular learning in neural networks. In *Proceedings of the sixth National conference on Artificial Intelligence-Volume 1*, pp. 279–284, 1987.

Adrien Bardes, Jean Ponce, and Yann LeCun. VicREG: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. doi:10.48550/arxiv.2105.04906.

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. MultiGrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019. doi:10.48550/arxiv.1902.05509.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "Siamese" time delay neural network. *Advances in Neural Information Processing Systems (NeurIPS)*, 6, 1993.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629. IEEE, 2021.

Hyesong Choi, Hyejin Park, Kwang Moo Yi, Sungmin Cha, and Dongbo Min. Salience-based adaptive masking: revisiting token dynamics for enhanced pre-training. In *European Conference on Computer Vision*, pp. 343–359. Springer, 2024.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 197–211, 2022.

Garrison W Cottrell, Paul Munro, and David Zipser. Learning internal representation from gray-scale images: An example of extensional programming. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 9, 1987.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18613–18624, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. In *European Conference on Computer Vision*, pp. 348–366. Springer, 2024.

Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10386–10395. IEEE, 2023.

Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, XuDong Wang, Adam Yala, Trevor Darrell, Alexei A Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=JT2KMuo2BV.

Anthony Fuller, James R Green, and Evan Shelhamer. Self-soupervision: Cooking model soups without labels. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=z0UM7y0L4r.

Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:40676–40693, 2023.

Maryam Haghighat, Peyman Moghadam, Shaheer Mohamed, and Piotr Koniusz. Pre-training with random orthogonal projection image modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=z4Hcegjzph.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Carlos Hinojosa, Shuming Liu, and Bernard Ghanem. ColorMAE: Exploring data-independent masking strategies in masked autoencoders. In *European Conference on Computer Vision*, pp. 432–449. Springer, 2024.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8126–8135. IEEE, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. *arXiv preprint arXiv:2304.01433*, 2023. doi:10.48550/arXiv.2304.01433.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2022.

Markus Karmann, Shile Li, Christian Internò, Bruno Andreis, David Klindt, Randall Balestriero, Jindong Gu, Philip Torr, Qi Zhang, Peng-Tao Jiang, Hao Zhang, Bo Li, and Onay Urfalioglu. MotionJEPA: Preventing temporal feature collapse by capturing visual changes in latent space. *arXiv preprint arXiv:2609.23881*, 2026. doi:10.48550/arxiv.2609.23881.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. ILIAS: Instance-level image retrieval at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14777–14787, 2025.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. SemMAE: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14290–14302, 2022a.

Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling MAE pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022b. doi:10.48550/arxiv.2205.10063.

Zhe Li, Zhongwen Rao, Lujia Pan, Pengyun Wang, and Zenglin Xu. Ti-MAE: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023. doi:10.48550/arxiv.2301.08871.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. PixMIM: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*, 2023a. doi:10.48550/arxiv.2303.02416.

Zhengqi Liu, Jie Gui, and Hao Luo. Good helper is around you: Attention-driven masked image modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1799–1807, 2023b.

Scott C Lowe, Anthony Fuller, Sageev Oore, Evan Shelhamer, and Graham W Taylor. Self-distillation of hidden layers for masked self-supervised representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pp. 604–621. Springer, 2022.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Lukas Picek, Klara Janouskova, Vojtech Cermak, and Jiri Matas. FungiTastic: A multi-modal dataset and benchmark for image categorization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 2046–2056, June 2025.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2018. doi:10.48550/arxiv.1704.00675.

F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018.

Jathushan Rajasegaran, Xinlei Chen, Rulilong Li, Christoph Feichtenhofer, Jitendra Malik, and Shiry Ginosar. Gaussian masked autoencoders. *arXiv preprint arXiv:2501.03229*, 2025. doi:10.48550/arxiv.2501.03229.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2021. URL https://openreview.net/forum?id=Zkj_VcZ6ol.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610. IEEE, 2007.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pp. 20026–20040. PMLR, 2022.

Jeongwoo Shin, Inseo Lee, Junho Lee, and Joonseok Lee. Self-guided masked autoencoder. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:58929–58954, 2024.

Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5461–5471. IEEE, 2023.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Computer Vision and Pattern Recognition*, 2021.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10375–10385. IEEE, 2023a.

Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2122–2131. IEEE, 2023b.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14648–14658. IEEE, 2022.

Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:3502–3516, 2022.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17923–17934. IEEE, 2023.

Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via coding rate regularization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=shTZSlk0HQ.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9-umxtNPx5E.

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Saining Xie, Hengshuang Zhao, Kaiming He, and Hu Xu. In pursuit of pixel supervision for visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31974–31984, 2026.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. doi:10.48550/arxiv.1910.04867.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. doi:10.48550/arxiv.2111.07832.

A Appendix

A.1 Additional pre-training details

We implement pre-training in JAX and train on TPUs. Each step samples 1,024 ImageNet-1K training images and makes two views of each (2,048 views per step), so our MAE baseline runs the same two-view pipeline, with each decoder receiving its own view's `CLS` token. We augment each view independently with a random resized crop (scale $[0.08, 1]$, bicubic), a random horizontal flip, and RandAugment (Cubuk et al., 2020), where weak/medium/strong apply 1/2/4 operations at magnitude 5/9/15 (magnitude noise std 0.5); we use no color jitter or random erasing. We then mask each view independently (ρ=75% unless stated otherwise). For color masking, we generate noise on the 14×14 patch grid for each view, filtering white Gaussian noise with a Gaussian (σ=2 patches) for red or a difference of Gaussians (σ=1 and 4 patches) for green, and we keep the lowest-scoring patches. The encoder and decoder use fixed 2D sin-cos position embeddings and no dropout or stochastic depth. The decoder has 8 blocks with head dimension 32 and width 128/256/512/512 for ViT-T/S/B/L encoders, where ViT-L has 24 blocks of width 1024 with 16 heads. We use AdamW (β_1=0.9, β_2=0.95, weight decay 0.05 on weight matrices only) with a peak learning rate of 6×10^{-4} ($1.5\times10^{-4} \times 1024/256$). The learning rate warms up linearly from 10^{-6} over 10/20/40 epochs for 100/200/800-epoch runs, then follows a cosine decay to 10^{-5}; we use no gradient clipping.

A.2 Additional fine-tuning details

We fine-tune on IN-1K following the ViT-B recipe of He et al. (2022). We train for 100 epochs with a batch size of 1,024, using AdamW (β_1=0.9, β_2=0.999, weight decay 0.05 on weight matrices only), layer-wise learning-rate decay of 0.65, and drop path of 0.1. The learning rate warms up linearly from 0 over 5 epochs, then follows a cosine decay to 10^{-6}. For augmentation, we use a random resized crop (scale $[0.08, 1]$), a horizontal flip, RandAugment (Cubuk et al., 2020) (2 operations, magnitude 9), random erasing (p=0.25), and either Mixup (α=0.8) or CutMix (α=1.0), chosen per batch with equal probability, with label smoothing of 0.1. For each pre-trained model, we sweep the base learning rate over $\{2.5, 5, 10, 20\}\times10^{-4}$ (peak learning rate = base $\times\, 1024/256$) and report the best. We evaluate on the 50K IN-1K validation images, resized to 256 pixels and center-cropped to 224×224.

A.3 Reference models

All reference models are ViT-B/16 pre-trained on IN-1K at 224×224 pixels. We download their public weights and evaluate them with the same frozen-feature protocols as our models. For MAE (He et al., 2022), we use the official 1600-epoch checkpoint, loaded through timm as `vit_base_patch16_224.mae`. For ColorMAE (Hinojosa et al., 2024), we use the official green-noise checkpoints pre-trained for 300, 800, and 1600 epochs (`https://huggingface.co/carlosh93/colormae`). LeJEPA (Balestriero & LeCun, 2025) has no official ViT-B/16 release, so we use the 100-epoch checkpoint of a public reproduction (`https://huggingface.co/OK-AI/lejepa-vitb16-pretrain-in1k`). For SimDINOv2 (Wu et al., 2025), we use the official 100-epoch checkpoint (`https://github.com/RobinWu218/SimDINO`). For Bootleg (Lowe et al., 2026), we use the 600-epoch checkpoint shared with us by the authors directly. We evaluate its EMA target encoder and discard its 4 register tokens before pooling. For the comparisons in Sec. 5, we use two further sets of official checkpoints. The first is the IN-1K-pre-trained ViT-S/16 CropMAE checkpoint (`https://github.com/alexandre-eymael/CropMAE`). The second is the ViT-B CroCo and CroCo v2 checkpoints (`https://github.com/naver/croco`); these pre-train on image pairs of 3D scenes rather than IN-1K.

A.4 Frozen-feature evaluation details

We resize each image's shorter side to 224 pixels, center-crop to 224×224, and extract the final-layer `CLS` token and the mean of the final-layer patch tokens. For kNN, we standardize each feature dimension using training-set statistics, L2-normalize, and weight each of the k nearest training neighbors' votes by $\exp(s/T)$, where s is the cosine similarity. We sweep k and T jointly, starting from a log-spaced 12×12 grid ($k\in[1, 200]$, $T\in[0.005, 0.5]$). Whenever the best value falls on an edge

of the grid, we extend that axis until the best (k, T) is interior. We cap k at 1,000. We select (k, T) on the evaluation split, identically for every model. For instance retrieval, animal re-identification, and Omniglot, we rank by cosine similarity between L2-normalized features, with no tuning. For ROxford5k, RParis6k, and ILIAS, we crop each query to its annotated bounding box; for Omniglot, we assign each test image the class of its nearest support image.

A.5 ALL RESULTS

For completion, we present results tables that include all 81 pre-trained models from Sec. 4.1 (located on the next two pages). We also present results of evaluations that we run using official/public checkpoints (located below).

			IN-1K		IN-21K 10-shot		m-VTAB Natural		m-VTAB Specialized		m-VTAB Structured		Instance Retrieval mAP							
			kNN		kNN		kNN		kNN		kNN		ROxford		RParis		ILIAS		SOP	
Model	Ep.	Pre-train GFLOPs	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool
Bootleg	600	4.15	72.6	64.8	15.9	12.3	61.9	56.6	77.2	78.0	27.2	27.9	13.3	9.3	40.5	27.7	10.7	5.0	34.9	33.2
LeJEPA	100	2.01	61.1	51.9	12.1	9.1	57.6	52.7	79.5	78.8	28.4	30.1	12.9	13.8	34.1	32.6	7.8	5.7	33.7	32.6
ColorMAE	300	1.08	42.7	45.7	6.3	7.3	45.2	49.9	74.9	74.0	29.5	28.7	8.9	6.3	23.1	15.1	2.9	2.1	30.1	20.9
ColorMAE	800	2.87	47.3	51.9	6.7	8.5	45.2	52.6	74.9	73.8	29.4	29.1	5.9	5.4	20.9	13.9	2.6	1.9	29.1	19.2
ColorMAE	1600	5.74	53.8	55.1	8.0	9.2	47.8	53.8	75.4	74.3	31.0	29.7	7.2	5.3	20.9	13.1	3.4	1.8	29.6	19.2
MAE	1600	5.74	48.1	48.0	7.0	7.2	44.7	47.0	73.6	72.5	29.5	28.1	3.8	4.8	11.3	10.9	1.9	1.5	18.7	16.8
SimDINOv2	100	2.91	78.1	63.8	21.1	12.2	71.8	60.7	80.8	79.1	28.6	30.8	26.6	18.1	62.2	45.6	25.3	9.4	33.3	33.4

			WildlifeReID 10K		iNat2021 mini		Omniglot		FungiTastic		State Information Prediction NMSE↓ $\times 100$							
			mAP		kNN		Acc.		kNN		Pong		Dino		Golf	
Model	Ep.	Pre-train GFLOPs	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool
Bootleg	600	4.15	18.0	16.7	18.7	11.9	42.0	45.8	31.4	27.0	0.8	3.2	8.5	12.5	6.1	6.8
LeJEPA	100	2.01	19.8	18.8	13.7	9.5	43.3	46.5	29.2	24.4	2.6	1.4	5.7	5.2	12.5	10.2
ColorMAE	300	1.08	15.0	12.6	6.9	6.4	54.8	43.0	17.2	16.7	2.1	20.1	3.2	5.4	5.9	6.0
ColorMAE	800	2.87	14.3	11.8	7.6	7.3	48.3	40.0	17.3	17.5	1.1	2.1	10.5	3.3	3.8	8.0
ColorMAE	1600	5.74	14.7	11.7	9.4	7.9	51.0	45.0	18.7	18.2	13.0	5.2	3.9	4.4	4.3	19.2
MAE	1600	5.74	10.9	10.8	6.9	6.1	37.3	39.8	16.0	15.5	3.4	10.2	8.3	3.5	28.3	12.3
SimDINOv2	100	2.91	23.2	19.8	33.9	14.6	66.0	66.3	39.7	27.8	1.7	0.6	5.3	4.4	7.5	5.9

				IN-1K					IN-21K 10-shot		m-VTAB Natural		m-VTAB Specialized		m-VTAB Structured		Instance Retrieval mAP							
				kNN		LP			kNN		kNN		kNN		kNN		ROxford		RParis		ILIAS		SOP	
Swing	Ep.	Aug. Strength	Color Mask	CLS	Pool	CLS	Pool	FT	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool
None	100	Weak	None	10.4	24.5	24.9	40.8	81.8	0.8	3.1	18.7	31.3	66.8	70.8	23.5	25.7	2.5	6.8	9.2	16.8	0.9	2.2	12.6	21.8
			Red	33.9	37.6	49.7	51.9	80.8	4.8	5.8	40.2	44.7	74.6	75.7	28.4	28.6	5.6	6.8	19.8	19.4	4.1	3.4	31.7	26.5
			Green	40.9	38.7	56.6	54.4	81.5	6.3	6.1	46.9	46.0	74.9	75.3	29.0	27.9	9.5	7.8	29.5	19.3	4.3	3.0	26.8	24.9
		Medium	None	33.4	34.3	49.2	48.5	80.9	4.6	4.9	39.0	40.3	74.7	74.9	27.5	27.2	10.4	6.6	26.3	17.9	3.4	2.5	27.4	23.6
			Red	29.1	32.3	44.1	46.0	80.3	4.0	4.7	36.8	41.4	72.8	74.3	26.4	26.1	6.1	5.9	18.0	17.4	3.6	2.8	30.8	24.9
			Green	36.0	36.0	51.4	51.4	80.8	5.3	5.4	43.6	43.7	73.5	75.6	29.4	27.8	9.7	6.5	27.0	18.3	4.0	2.7	25.5	24.3
		Strong	None	28.2	31.3	42.6	44.1	80.3	3.7	4.3	35.5	38.0	73.3	74.2	26.4	26.5	7.5	5.8	22.5	17.1	3.1	2.4	27.8	23.4
			Red	20.9	27.4	35.1	39.6	79.5	2.5	3.8	29.3	37.5	72.8	74.4	25.2	26.3	4.9	5.4	13.8	14.5	3.1	2.4	29.1	23.3
			Green	26.7	32.7	43.1	46.2	80.0	3.5	4.7	33.4	40.0	74.2	74.2	26.3	27.6	4.6	6.0	16.6	17.2	3.1	2.5	32.0	23.6
	200	Weak	None	39.8	40.7	56.8	55.7	82.3	5.7	6.2	43.9	45.3	74.4	74.0	29.2	27.6	8.7	6.6	25.0	16.5	3.8	2.5	27.4	23.1
			Red	41.5	42.8	56.9	58.0	81.9	6.5	7.1	45.7	48.8	76.0	76.1	27.6	28.5	6.7	7.3	20.3	20.4	4.4	3.7	33.5	27.0
			Green	45.2	43.8	60.6	59.5	82.3	7.2	7.0	49.4	48.8	74.6	75.2	30.4	29.2	12.5	7.0	29.1	17.3	4.1	2.9	25.3	24.0
		Medium	None	35.3	38.1	53.5	53.0	81.3	5.0	5.7	41.4	43.3	75.2	75.0	29.1	27.1	8.7	6.6	22.9	16.7	3.5	2.3	26.4	23.0
			Red	34.4	38.0	50.2	53.3	81.1	5.0	6.0	40.7	45.1	75.5	76.2	27.7	28.2	5.5	6.3	18.4	18.2	4.1	3.5	34.1	26.8
			Green	41.7	41.3	57.2	57.0	81.8	6.6	6.6	47.9	47.5	74.1	75.6	29.4	28.9	11.3	6.8	30.2	16.9	4.1	2.8	25.2	24.2
		Strong	None	33.8	35.3	49.5	49.3	80.9	4.6	5.0	39.6	40.6	72.4	75.0	28.1	26.4	8.0	5.6	25.5	16.2	3.6	2.2	27.3	22.4
			Red	26.3	33.2	42.4	47.3	80.3	3.4	4.9	35.1	41.5	73.5	75.3	25.6	27.4	4.7	5.0	16.7	16.1	3.3	2.8	32.3	25.0
			Green	33.4	37.4	51.0	52.6	80.8	4.7	5.7	39.2	44.1	74.7	74.9	28.6	28.3	5.5	6.3	18.7	17.5	3.5	2.9	32.7	24.1
	800	Weak	None	37.9	46.4	60.7	61.6	82.8	4.8	7.2	39.5	48.3	72.6	73.8	28.5	28.6	7.0	5.5	17.8	12.4	2.9	2.1	25.9	20.1
			Red	49.4	52.8	64.8	65.9	82.6	7.8	9.6	49.6	54.0	76.9	76.8	29.5	29.8	5.9	6.9	19.5	18.8	4.3	3.3	36.7	25.9
			Green	53.3	52.8	67.4	67.3	83.0	8.9	9.0	51.9	53.0	76.0	74.8	31.0	30.0	9.0	6.0	25.3	15.3	3.6	2.5	27.3	22.1
		Medium	None	38.6	43.3	58.1	58.6	82.2	5.3	6.6	40.0	44.9	75.2	74.8	27.4	27.4	6.9	5.8	19.5	13.6	2.8	2.1	26.9	20.6
			Red	41.1	48.0	59.5	62.6	81.9	5.9	8.4	45.1	51.5	77.0	77.6	28.5	30.2	5.0	6.5	15.0	17.8	3.6	3.1	35.5	25.8
			Green	51.8	49.5	65.0	64.6	82.4	8.8	8.5	55.2	53.0	74.1	75.0	30.2	29.1	11.1	6.0	30.1	14.9	4.0	2.4	25.5	22.3
		Strong	None	36.1	40.7	54.9	55.9	81.8	5.1	6.1	39.2	44.3	75.2	75.8	28.8	28.4	5.2	5.4	18.7	13.2	3.3	2.2	26.4	20.6
			Red	32.6	42.4	51.5	57.2	81.0	4.6	6.9	37.1	47.7	76.8	76.9	27.1	28.7	4.4	5.6	13.9	16.4	3.6	2.9	37.1	25.5
			Green	39.0	45.5	58.2	60.8	81.7	5.7	7.4	42.7	49.3	76.5	76.6	28.9	29.4	5.6	6.1	14.2	14.5	3.0	2.6	29.4	22.3
Half	100	Weak	None	45.9	35.7	56.7	51.8	81.9	8.3	5.2	51.5	42.3	75.3	74.2	30.2	27.6	12.4	7.6	36.9	18.1	6.9	2.7	28.0	24.4
			Red	37.0	36.8	48.6	51.1	81.0	6.2	5.6	44.9	44.4	75.0	75.2	29.4	28.1	8.9	7.2	26.4	20.6	4.2	3.4	29.8	27.7
			Green	44.2	38.1	54.9	53.6	81.6	8.0	5.9	50.5	45.3	76.3	75.0	29.8	29.0	10.9	7.7	33.4	19.5	5.1	3.0	28.8	25.2
		Medium	None	42.4	33.9	53.9	48.6	80.8	7.1	4.9	46.7	40.8	74.8	74.9	30.0	27.4	13.9	7.3	37.2	17.8	6.1	2.5	28.5	23.7
			Red	33.3	32.1	44.2	46.3	80.3	5.4	4.7	41.9	41.2	74.9	74.6	28.9	27.9	8.7	6.4	25.0	17.4	4.2	3.0	30.4	26.1
			Green	42.1	35.6	52.6	51.1	80.9	7.5	5.4	49.0	43.0	75.6	74.4	30.4	28.6	11.2	7.2	34.4	19.2	5.1	3.1	27.7	24.9
		Strong	None	36.9	31.7	49.1	44.3	80.1	5.9	4.3	43.1	38.7	75.1	74.7	28.5	27.0	12.7	5.8	36.5	16.4	5.1	2.4	27.5	23.1
			Red	26.8	27.2	38.2	39.7	79.5	4.1	3.8	35.9	37.4	74.8	74.1	27.4	26.7	7.8	5.6	21.1	15.1	3.4	2.6	29.3	23.8
			Green	38.0	32.7	49.0	46.4	79.9	6.5	4.8	45.2	40.4	75.8	74.8	29.7	28.3	10.4	6.4	32.9	17.6	4.6	2.6	28.8	24.0
	200	Weak	None	46.2	38.8	56.7	55.0	82.2	8.1	5.9	52.2	44.2	75.0	74.3	31.8	28.0	12.2	7.1	36.5	17.1	6.1	2.5	26.3	23.7
			Red	41.7	41.8	54.1	57.0	81.6	7.3	6.8	48.6	47.5	76.0	76.2	29.3	28.2	9.3	6.8	28.2	20.3	4.7	3.5	30.9	28.3
			Green	47.1	41.7	58.5	58.3	82.4	8.7	6.7	53.2	47.6	75.7	74.8	30.6	28.2	10.9	7.1	33.2	17.6	5.2	3.0	28.9	25.2
		Medium	None	50.3	37.8	59.7	53.2	81.5	9.3	5.6	53.9	43.7	76.4	75.3	31.3	28.0	14.8	6.6	42.2	16.6	7.8	2.3	28.9	22.7
			Red	39.5	38.2	50.7	53.4	81.2	6.8	6.0	46.3	45.5	77.1	76.6	29.2	27.9	8.8	6.4	28.2	18.9	4.7	3.3	31.7	27.8
			Green	46.4	39.9	56.9	56.2	81.6	8.5	6.3	52.8	47.0	77.8	75.9	30.6	28.2	11.7	7.6	36.3	18.8	5.6	3.2	28.1	25.3
		Strong	None	44.5	35.6	55.1	49.0	80.7	7.7	5.1	48.8	41.2	75.6	74.8	29.5	27.7	13.0	6.1	39.4	15.6	6.8	2.2	28.8	22.2
			Red	35.1	33.3	46.5	47.6	80.4	5.7	4.9	42.4	41.8	76.7	74.9	28.8	28.2	8.7	5.5	24.6	16.5	4.7	2.8	32.8	25.5
			Green	44.8	37.4	54.8	52.3	80.9	8.1	5.7	50.6	44.0	77.5	75.4	30.4	29.0	11.3	6.4	35.7	16.8	5.6	2.8	29.5	24.2
	800	Weak	None	33.1	40.7	52.7	57.6	82.7	4.7	6.1	44.8	44.7	68.0	74.9	30.2	28.0	10.1	5.8	23.8	12.9	2.9	2.0	17.2	21.5
			Red	44.5	48.3	59.6	63.1	82.5	7.5	8.4	50.3	51.5	73.7	76.7	30.9	29.3	8.8	5.7	25.4	16.0	4.3	2.6	31.3	27.1
			Green	38.3	46.6	58.0	62.3	83.1	5.7	7.6	46.8	49.4	69.5	75.9	30.4	29.9	9.8	6.0	25.5	16.0	3.2	2.4	21.4	23.7
		Medium	None	56.1	42.7	63.9	59.0	82.3	11.0	6.5	57.2	46.3	77.2	74.9	32.0	28.8	18.3	5.9	44.1	13.4	9.0	2.0	31.0	20.9
			Red	44.8	45.4	57.7	60.5	81.9	7.9	7.8	50.7	49.9	77.0	77.3	30.4	29.5	9.2	5.7	27.9	15.8	4.9	3.0	32.6	27.9
			Green	41.3	44.0	57.7	60.8	82.5	6.5	7.1	49.6	48.6	71.3	76.0	30.2	29.7	10.8	6.0	29.5	14.2	3.8	2.2	23.7	23.1
		Strong	None	54.1	40.3	62.3	56.2	82.0	10.3	6.1	55.6	45.2	79.0	75.5	30.6	28.4	15.1	5.6	41.1	13.1	7.2	2.0	30.6	20.3
			Red	45.6	42.0	56.4	57.0	81.2	8.3	6.9	51.0	48.4	78.9	77.3	30.0	29.2	9.3	5.5	29.2	15.7	5.6	3.0	35.4	27.0
			Green	47.7	43.1	59.0	59.0	81.9	8.7	7.1	53.3	48.2	76.0	76.1	29.6	28.7	11.4	6.0	34.7	14.9	4.7	2.3	27.4	23.1
Full	100	Weak	None	39.1	33.7	49.6	48.7	81.7	6.7	5.0	48.6	41.4	73.7	74.8	30.8	28.3	10.7	7.7	32.4	19.9	4.2	2.7	24.5	25.1
			Red	34.5	34.1	45.2	48.2	81.3	5.8	5.1	44.8	42.4	74.6	75.1	28.8	28.4	9.1	6.9	26.0	20.0	3.8	3.2	28.0	28.6
			Green	37.2	35.2	47.8	50.3	81.9	6.2	5.4	47.1	42.6	72.1	75.2	29.3	29.1	10.1	7.8	30.4	20.6	3.7	2.9	24.1	26.2
		Medium	None	43.1	33.3	51.2	47.5	81.0	7.7	4.8	50.4	41.2	75.7	74.6	30.9	26.9	12.6	7.5	36.3	19.2	5.6	2.6	25.8	24.3
			Red	34.5	32.1	43.1	45.6	80.3	5.8	4.7	43.7	41.9	75.1	75.3	29.2	27.4	8.9	7.2	26.8	19.0	4.1	3.2	30.1	27.2
			Green	40.5	34.8	48.8	49.7	81.1	7.1	5.2	49.4	43.3	76.0	75.1	29.2	29.0	10.7	7.2	32.9	19.9	4.7	2.9	27.4	25.0
		Strong	None	40.8	30.3	48.2	43.3	80.4	6.9	4.2	40.1	35.5	75.1	74.0	28.4	26.1	10.8	6.2	33.8	16.6	5.0	2.4	30.2	23.0
			Red	30.4	28.0	39.5	40.5	79.8	4.9	4.0	39.2	38.4	75.8	74.7	27.7	26.8	8.2	6.4	24.2	16.7	4.2	2.9	29.9	24.9
			Green	39.8	32.0	48.2	46.6	80.2	7.1	4.7	47.1	40.1	76.4	75.5	29.9	28.0	10.5	6.9	33.1	18.5	4.8	2.7	28.9	24.0
	200	Weak	None	28.5	33.9	46.1	49.7	82.4	4.1	5.1	42.5	41.5	66.4	74.7	30.7	28.6	11.4	6.9	23.8	17.3	2.7	2.3	17.9	24.4
			Red	36.7	36.8	48.1	51.6	81.6	6.1	5.7	45.9	44.2	74.6	76.0	29.6	28.4	9.4	6.7	25.6	19.9	4.0	3.2	29.1	28.6
			Green	35.2	37.1	46.9	52.8	82.5	5.3	5.7	44.9	43.4	69.3	75.5	30.1	28.2	11.2	7.0	25.8	18.8	3.6	2.8	22.5	26.7
		Medium	None	39.1	35.6	51.5	50.6	81.5	6.5	5.3	48.4	42.3	73.3	75.3	30.3	28.3	11.5	6.8	33.1	18.1	4.2	2.5	22.0	24.3
			Red	37.4	36.2	47.9	50.7	81.0	6.4	5.6	46.6	43.8	75.2	76.2	29.6	28.2	9.3	6.8	27.6	19.0	4.2	3.1	29.0	28.3
			Green	37.8	36.6	49.0	52.0	81.7	6.1	5.6	47.9	44.1	73.8	75.9	30.2	28.7	9.8	7.2	29.7	18.9	3.9	2.9	24.1	25.9
		Strong	None	45.8	35.1	52.9	48.8	80.9	8.4	5.2	52.4	41.9	76.4	75.7	30.7	27.6	13.5	6.2	39.2	16.9	5.9	2.4	25.9	23.6
			Red	37.1	33.4	46.1	47.7	80.2	6.4	5.0	44.5	41.8	76.8	75.8	29.3	28.5	8.8	6.1	27.4	17.7	4.8	3.1	32.5	27.1
			Green	42.7	36.6	50.4	51.2	81.0	7.6	5.7	50.0	44.7	76.7	75.9	29.5	29.3	11.1	6.9	33.9	18.5	5.1	2.9	28.4	25.5
	800	Weak	None	22.9	33.2	37.3	50.6	82.8	3.0	5.0	36.8	40.8	60.9	72.8	29.7	29.0	7.0	5.4	16.7	13.2	1.8	2.1	13.5	22.3
			Red	36.4	42.5	52.2	58.2	82.5	5.6	6.8	44.9	47.0	70.0	76.6	31.4	29.8	8.2	5.6	22.0	14.9	3.5	2.4	27.3	27.2
			Green	29.4	39.7	44.2	55.8	83.1	3.9	6.1	40.2	44.4	64.4	75.0	30.6	29.7	7.8	5.7	20.6	15.4	2.9	2.5	19.0	26.3
		Medium	None	24.6	34.2	41.8	50.7	82.3	3.2	5.0	38.5	41.1	62.8	75.0	30.4	28.9	7.1	5.6	19.7	14.1	1.8	2.1	13.8	23.0
			Red	37.9	40.0	52.6	56.5	82.0	6.1	6.4	46.8	45.4	72.8	77.4	30.7	29.2	9.5	5.6	23.5	14.9	3.9	2.5	28.5	27.8
			Green	31.7	38.3	45.7	54.9	82.5	4.5	5.9	42.5	44.7	66.6	75.6	30.4	29.4	8.8	6.0	23.2	15.7	2.8	2.3	19.8	25.8
		Strong	None	34.5	36.3	48.4	51.8	82.2	5.1	5.5	46.0	42.9	69.9	76.7	30.7	28.7	11.5	5.7	25.6	14.1	3.0	2.1	17.7	22.7
			Red	41.2	38.7	51.7	54.2	81.3	7.1	6.2	49.0	45.7	76.4	77.7	29.3	28.7	9.4	6.1	26.9	17.0	4.8	3.1	33.1	29.1
			Green	34.5	39.0	49.5	55.1	81.8	5.1	6.1	44.0	45.4	69.8	76.9	30.5	29.0	9.5	6.0	24.7	15.1	3.1	2.3	20.4	24.8

				WildlifeReID 10K		iNat2021 mini		Omniglot		FungiTastic		State Information Prediction NMSE×100↓					
				mAP		kNN		Acc.		kNN		Pong		Dino		Golf	
Swing	Ep.	Aug. Strength	Color Mask	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool	CLS	Pool
None	100	Weak	None	9.4	12.5	0.8	3.0	11.3	21.0	5.8	10.0	2.6	1.2	9.2	4.5	14.1	15.2
			Red	15.9	14.5	5.8	5.6	39.5	40.0	16.5	17.4	3.3	1.4	13.6	10.8	12.0	7.5
			Green	14.7	14.2	7.8	5.4	48.3	44.8	18.2	16.5	3.6	0.7	3.7	5.3	9.3	5.8
		Medium	None	14.6	13.5	5.3	5.0	36.8	38.5	15.1	15.6	242.8	5.6	23.7	6.1	98.4	14.6
			Red	15.2	13.6	4.2	4.6	37.0	41.5	13.3	14.7	7.6	2.8	9.6	4.5	8.0	7.8
			Green	13.9	14.0	5.9	5.2	41.5	44.5	16.7	16.1	40.3	1.1	8.6	5.2	17.1	7.9
		Strong	None	14.1	13.2	4.0	4.6	32.3	32.5	13.2	14.9	80.6	8.3	33.0	4.8	36.4	19.8
			Red	13.0	12.8	3.0	3.7	21.3	30.0	10.2	13.2	44.0	4.4	11.7	11.3	11.6	19.5
			Green	14.9	13.3	4.3	4.8	32.0	40.8	14.5	15.4	8.7	14.4	9.3	8.8	14.6	9.3
	200	Weak	None	14.5	13.5	6.6	6.0	49.8	39.0	15.5	15.6	103.0	7.3	62.1	5.7	60.9	13.7
			Red	16.2	14.6	7.3	6.8	42.0	40.3	17.9	18.7	1.4	5.9	4.9	6.6	7.2	14.4
			Green	14.6	13.9	8.4	6.3	46.0	45.5	18.4	17.2	77.7	7.2	26.9	3.1	85.2	6.7
		Medium	None	13.6	13.2	6.2	5.9	41.5	38.3	15.6	16.2	63.9	3.5	31.6	5.4	61.4	12.7
			Red	15.6	14.3	5.7	5.8	40.8	42.3	16.4	17.7	3.5	16.4	7.8	5.0	8.7	7.0
			Green	14.2	13.9	7.7	6.3	42.0	47.3	18.1	17.4	61.6	1.4	41.8	4.4	13.0	6.6
		Strong	None	14.1	12.7	5.6	5.4	38.8	34.8	15.1	16.0	10.9	6.9	18.3	7.3	47.8	14.2
			Red	13.0	13.3	4.0	5.1	25.5	39.3	12.6	16.2	4.5	2.0	8.3	7.8	10.4	11.9
			Green	15.6	13.7	5.6	5.6	36.8	43.3	16.9	16.7	23.9	1.5	9.5	4.9	7.5	10.0
	800	Weak	None	13.1	12.1	5.7	7.0	43.5	45.5	14.2	16.6	65.5	5.6	25.9	4.4	65.1	10.1
			Red	15.5	14.4	9.1	8.9	43.0	38.5	20.5	22.1	5.4	9.1	8.2	12.6	4.2	14.6
			Green	15.0	13.2	9.8	7.5	47.5	44.5	19.6	18.5	25.1	1.3	7.6	3.9	41.7	8.0
		Medium	None	13.6	12.2	6.2	6.8	51.0	41.0	15.4	16.8	192.5	16.8	22.6	4.8	82.7	10.2
			Red	14.1	14.2	7.0	8.1	39.8	44.5	18.2	21.0	1.7	3.6	15.7	8.1	5.9	29.0
			Green	14.0	13.3	10.5	7.8	47.5	46.5	20.5	18.8	15.2	1.8	9.5	5.4	42.4	4.7
		Strong	None	13.9	12.1	5.6	6.7	49.0	44.0	15.0	17.1	4.0	32.5	6.0	3.8	5.8	12.2
			Red	15.6	13.5	5.5	7.3	35.8	43.5	16.4	19.7	7.0	8.1	13.5	5.7	11.5	9.3
			Green	13.8	13.0	6.6	7.3	38.8	45.0	18.5	18.9	7.6	9.2	16.6	4.1	3.3	7.9
Half	100	Weak	None	18.5	14.0	10.8	5.2	71.8	42.8	20.2	14.9	0.6	2.8	2.5	5.4	8.2	11.7
			Red	17.5	15.1	6.8	5.5	62.8	48.0	18.4	16.9	4.1	1.5	3.6	9.8	4.2	5.8
			Green	18.3	14.5	9.8	5.4	71.3	50.0	21.3	16.2	1.1	0.8	2.5	5.2	4.4	5.7
		Medium	None	16.4	13.5	8.6	5.1	46.3	41.8	18.3	15.0	3.4	3.8	4.6	4.1	9.4	11.4
			Red	17.4	14.0	6.1	4.6	56.0	44.3	17.4	15.2	2.7	2.1	6.0	5.2	7.2	6.2
			Green	17.3	14.3	8.9	5.1	71.8	50.0	20.1	15.8	0.8	3.4	2.7	5.8	4.8	6.5
		Strong	None	15.8	13.1	7.2	4.8	41.3	29.5	17.4	15.0	2.4	23.2	8.5	6.8	13.0	25.2
			Red	16.5	13.0	4.3	3.7	45.8	35.0	14.3	12.8	2.1	6.4	7.4	15.9	8.3	11.9
			Green	16.8	13.6	7.8	4.9	58.8	40.5	19.2	15.6	0.8	6.3	3.5	5.3	4.7	7.8
	200	Weak	None	17.8	14.0	9.7	5.8	80.0	42.8	18.1	15.5	1.5	31.8	1.5	3.8	3.9	11.3
			Red	17.5	15.0	8.1	6.5	70.8	43.5	19.3	18.9	0.9	2.3	2.3	8.5	4.6	4.3
			Green	18.0	14.5	10.1	6.1	75.8	48.0	20.5	16.9	0.8	14.5	2.1	2.9	12.1	4.9
		Medium	None	17.8	13.2	13.1	6.0	68.0	37.5	22.9	16.3	1.1	10.0	4.1	3.1	6.2	9.6
			Red	18.0	14.6	7.8	6.0	63.0	46.5	19.9	18.0	1.0	4.0	4.3	6.8	4.6	9.5
			Green	18.0	14.6	10.5	6.0	82.0	53.0	21.2	16.8	0.9	6.4	2.4	4.2	3.8	4.4
		Strong	None	16.7	12.8	10.1	5.8	52.5	39.0	20.4	16.4	2.1	29.9	5.6	4.0	6.9	15.8
			Red	17.9	13.4	6.6	5.2	53.3	41.8	18.8	16.0	0.8	14.6	9.4	5.6	8.3	12.7
			Green	17.7	13.7	10.1	5.8	74.0	47.0	21.7	16.7	0.7	1.9	2.9	5.0	7.4	9.0
	800	Weak	None	13.0	12.6	5.1	6.0	85.8	47.0	10.7	14.9	0.7	1.5	1.5	2.7	3.6	5.8
			Red	16.9	14.3	7.8	8.0	81.8	42.0	18.0	20.5	0.9	15.2	1.6	16.4	2.4	3.1
			Green	14.4	13.6	5.5	6.7	83.0	41.5	13.0	17.2	0.5	2.3	1.3	3.5	2.9	2.9
		Medium	None	18.8	12.5	15.1	7.1	77.3	46.8	23.6	17.2	0.8	2.6	2.6	7.0	6.4	9.9
			Red	17.3	14.4	8.8	7.6	81.8	55.3	20.2	20.3	0.6	3.7	2.2	8.0	3.0	3.7
			Green	15.2	13.5	6.9	6.7	88.8	47.8	15.0	17.8	0.6	6.2	1.4	7.5	2.8	3.3
		Strong	None	17.5	12.2	14.4	6.6	64.5	43.3	24.9	17.0	0.7	2.0	4.1	5.0	5.1	9.0
			Red	18.9	14.0	10.1	7.1	73.0	54.0	23.1	19.2	3.3	31.9	3.4	12.3	2.9	4.8
			Green	17.0	13.2	10.2	7.2	84.0	56.0	19.7	18.4	0.7	10.3	2.5	4.5	3.7	4.3
Full	100	Weak	None	17.4	14.6	7.8	5.0	84.8	44.8	15.6	14.5	0.9	2.8	1.8	2.3	3.8	3.6
			Red	16.4	15.2	6.2	5.1	77.8	49.8	16.5	16.5	1.3	1.8	2.3	7.0	3.8	4.6
			Green	15.4	14.8	6.7	5.2	78.0	49.5	15.5	16.1	1.1	0.8	1.5	2.4	2.9	3.6
		Medium	None	18.6	14.0	9.7	5.0	79.0	45.8	18.7	14.9	0.9	4.9	2.5	3.0	4.2	9.4
			Red	17.8	14.7	6.6	4.7	68.5	48.8	18.0	15.4	0.7	3.5	3.4	5.1	3.9	7.1
			Green	18.3	14.5	8.6	5.4	80.5	53.3	18.8	16.1	1.0	1.8	2.2	3.0	4.2	5.7
		Strong	None	16.9	13.1	9.4	4.7	69.0	39.8	20.7	14.5	0.9	6.3	3.9	7.1	6.1	15.0
			Red	17.6	13.7	5.6	3.9	53.0	40.8	16.8	13.8	1.8	3.8	4.9	5.7	5.4	5.4
			Green	18.2	13.9	8.7	4.7	77.8	48.8	20.4	15.0	0.8	1.0	3.9	8.5	3.4	6.1
	200	Weak	None	13.5	13.9	4.2	5.1	86.8	43.0	9.6	14.5	0.8	4.8	1.3	2.3	3.7	3.6
			Red	16.6	15.0	6.7	5.7	80.0	45.0	17.2	17.6	0.6	9.0	1.3	12.1	2.6	4.5
			Green	15.1	14.6	5.6	5.4	87.3	47.3	13.3	16.3	0.7	41.1	1.3	3.0	2.8	8.8
		Medium	None	15.4	14.1	7.9	5.7	83.3	47.3	15.0	15.6	1.4	9.6	3.2	4.7	5.0	5.6
			Red	16.7	15.1	7.1	5.7	73.8	49.3	17.7	17.3	1.4	2.6	2.4	6.2	4.1	4.6
			Green	15.2	14.7	6.8	5.6	88.3	50.0	15.5	16.7	0.8	42.0	1.7	2.7	2.7	3.4
		Strong	None	18.6	13.6	11.4	5.8	79.0	45.3	20.5	16.1	0.9	6.6	3.6	3.8	4.3	10.0
			Red	18.6	14.4	7.4	5.2	62.8	47.3	19.7	16.1	1.9	10.9	4.1	4.6	3.7	12.4
			Green	18.3	14.4	9.4	5.9	84.3	51.8	19.9	17.2	0.7	1.4	2.5	7.4	3.2	6.7
	800	Weak	None	10.3	13.0	3.0	4.5	74.0	49.3	7.5	12.8	0.9	0.9	1.2	1.4	3.4	3.7
			Red	15.9	13.9	5.7	6.7	83.8	35.5	14.0	17.8	0.5	10.1	1.3	7.2	2.4	3.5
			Green	13.4	14.2	3.8	5.6	74.0	46.5	9.8	15.9	0.5	1.1	1.3	6.6	3.0	4.1
		Medium	None	10.8	13.1	3.3	5.0	77.0	47.3	8.0	13.7	0.5	18.0	1.5	6.0	3.5	3.3
			Red	16.4	14.2	6.6	6.4	88.3	44.8	15.9	18.4	0.5	1.9	1.4	6.5	2.5	5.6
			Green	13.7	14.1	4.6	5.7	87.3	49.0	11.4	16.5	0.5	1.3	1.4	4.3	3.2	3.0
		Strong	None	13.4	13.0	5.6	5.6	89.8	44.3	11.6	15.6	0.6	4.5	1.6	14.1	4.8	4.9
			Red	17.4	14.9	8.1	6.3	83.3	53.0	19.7	18.2	0.5	2.8	2.5	6.0	2.4	3.5
			Green	14.1	13.6	5.2	6.4	89.5	45.3	12.6	17.7	0.6	4.3	1.6	3.6	3.2	2.9