We would like to thank the reviewers for their careful and positive reviews. Due to space constraints, we respond to some of the reviewers' comments and questions here. All comments will be addressed in the final version of the paper.

Reviewers 2 and 4 suggest discussing additional related work, such as unsupervised pretraining and reducing the amount of pre-training data. We agree that these are relevant and will include them in the discussion of related work.

Reviewer 1 **Q:** "*I'm a bit confused about the experiment setting of Section 2.5 [...]*" **A:** In the setting of Tables 1-7, we had 3 conv layers, and one fully connected layer, but in general we can look at the first L conv layers in a neural network with $\geq L$ conv layers. All of the layers are computed from the input images (without using labels or learning). The simplest procedure (which already gives the main benefit, cf. Table 2) is this: The first layer uses the training input directly, computes the covariance matrix Σ_x of patches, and uses as k conv filters the eigenvectors e_i (normalized to length 1) corresponding to the k largest eigenvalues σ_i^2 of Σ_x. After computing this first layer, we can also compute the output of the first layer on the training images, which then becomes the input to the second layer. Then we compute the second layer in the same way using these transformed representations as input, and so on. Variants of this procedure include multiplying the filters e_i with $\tau_i = f(\sigma_i)$ for some assumed function $f(\sigma)$, or sampling the filters from a normal distribution $\mathcal{N}(0, \Sigma_w)$ where Σ_w has eigenvectors e_i with eigenvalues τ_i^2. These variants give very similar results (see Tables 1-4).

Reviewer 2 **Q:** "*It seems there will be a trade off between the bottom-up eigenspace alignment and the top-down task-specificity. The theoretical relationship between these two forces is left for future work.*" **A:** Thank you for this accurate summary. We agree with your assessment.

Q: "*The authors propose no broader societal or ethical impacts of their work. I think an argument can be made that work on understanding deep learning can contribute to interpretable and explainable AI, which has potential to help identify sources of bias, for example. I encourage the authors to try to write something more meaningful in that section.*" **A:** We will extend the discussion in the *Broader Impact* section following your advice.

Q: "*It is claimed that the present results help to explain critical learning stages in DNNs. This claim [...] is not justified clearly anywhere in the text.*" **A:** Our experiments suggest that critical learning periods that reduce learning capacity (as described in detail in Achille et al. [1]) occur partially because of the specialization that takes place at the upper layer. In Figures 7, 10 & 11, we show that neurons at the upper layer do specialize; meaning that a single neuron tends to be activated by a few images in the training sample only. This has a negative effect when the distribution of data changes. Figure 24 in Appendix F shows that neural activation at the upper layer can drop abruptly and permanently once the switch to the downstream task takes place. We will add an additional brief explanation in the paper.

Reviewer 3 **Q:** "*The paper provides some reasoning for the shape of $f(\sigma)$ which could be tested but it is not easy to follow—why is it reasonable to expect that the large eigenvalues will dominate the output of the layers or what does it mean for backprop to capture this signal?*" **A:** We find experimentally that e.g. for 5×5 patches on CIFAR10 $\sigma_1^2 \approx 12$ and $\sum_{i=2}^{75} \sigma_i^2 \approx 6$, so the first direction dominates the input (see first table in Appendix D.1). Using an orthonormal basis e_i of eigenvectors of Σ_x and Σ_w, we can decompose the variance of the output to a neuron $\mathbb{E}_x[\langle w, x\rangle^2]$ as $\sum_i \langle w, e_i\rangle^2 \mathbb{E}_x[\langle e_i, x\rangle^2] = \sum_i \langle w, e_i\rangle^2 \cdot \sigma_i^2$. Averaging over w gives $\sum_i \tau_i^2 \cdot \sigma_i^2$. So if $f(\sigma)$ would be increasing, direction e_1 would also dominate the output. We speculate that backprop finds a near optimal solution, and it seems likely that one component dominating is not optimal when there is also important information in the other components.

Q: "*The existence of positive and negative transfer has been mentioned a few times, but the precise cases in which they arise have not been specified.*" **A:** As Reviewer 2 mentioned, there is a tradeoff between both effects: the positive effect at the early layers and the negative effect at the later layers. Depending on the setting, such as initialization (cf. Figure 1) or the DNN architecture (cf. Figure 8), one effect may dominate the other.

Q: "*$f(\sigma)$ is not properly defined. The experimental setup for the synthetic case is not properly described.*" **A:** We will make the definition of $f(\sigma)$ (Equation (2)) and the synthetic experimental setup in Appendix E.1. more explicit.

Reviewer 4 **Q:** "*I would expect that with increasing kernel size of the convolutions, the correlation between patches increases and with that potentially the misalignment score. [...] This would also align with the specialization observation in Sec. 3.*" **A:** Yes, we believe that correlation effects can be observed. Some results in this direction are presented in Figure 22 in Appendix E.2, where we observe that "*Correlations between neighboring patches create deviation from the curve seen for the "ideal" case [...]. Smaller strides lead to stronger correlations and stronger deviation.*"

Q: "*With the findings in the paper it should be possible to construct a set of weights that aligns with the observed data distribution. [...]*" **A:** Yes, exactly. This is done in an example in Section 2.5. and we are considering this as an interesting future direction of research.