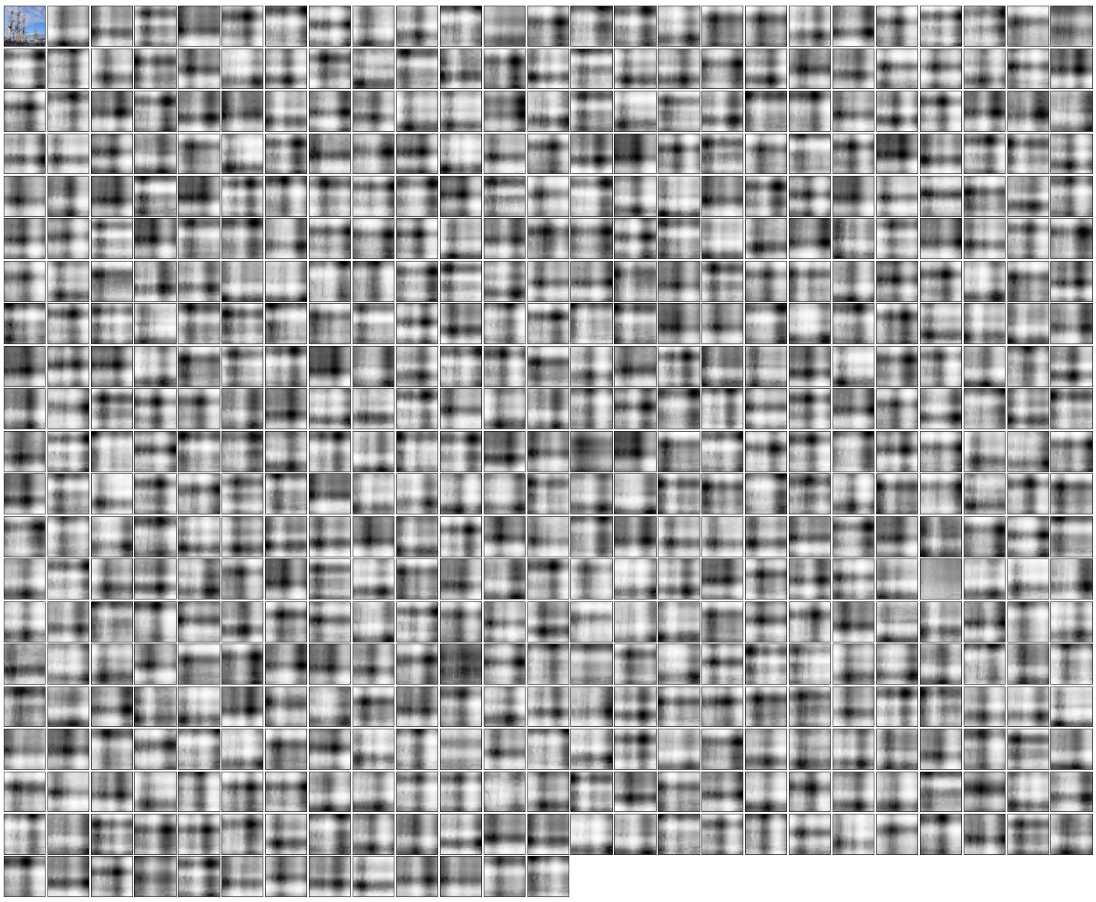

Figure 11. Example attention maps from the **first cross-attend** of an ImageNet network trained with **2D Fourier feature** position encodings.

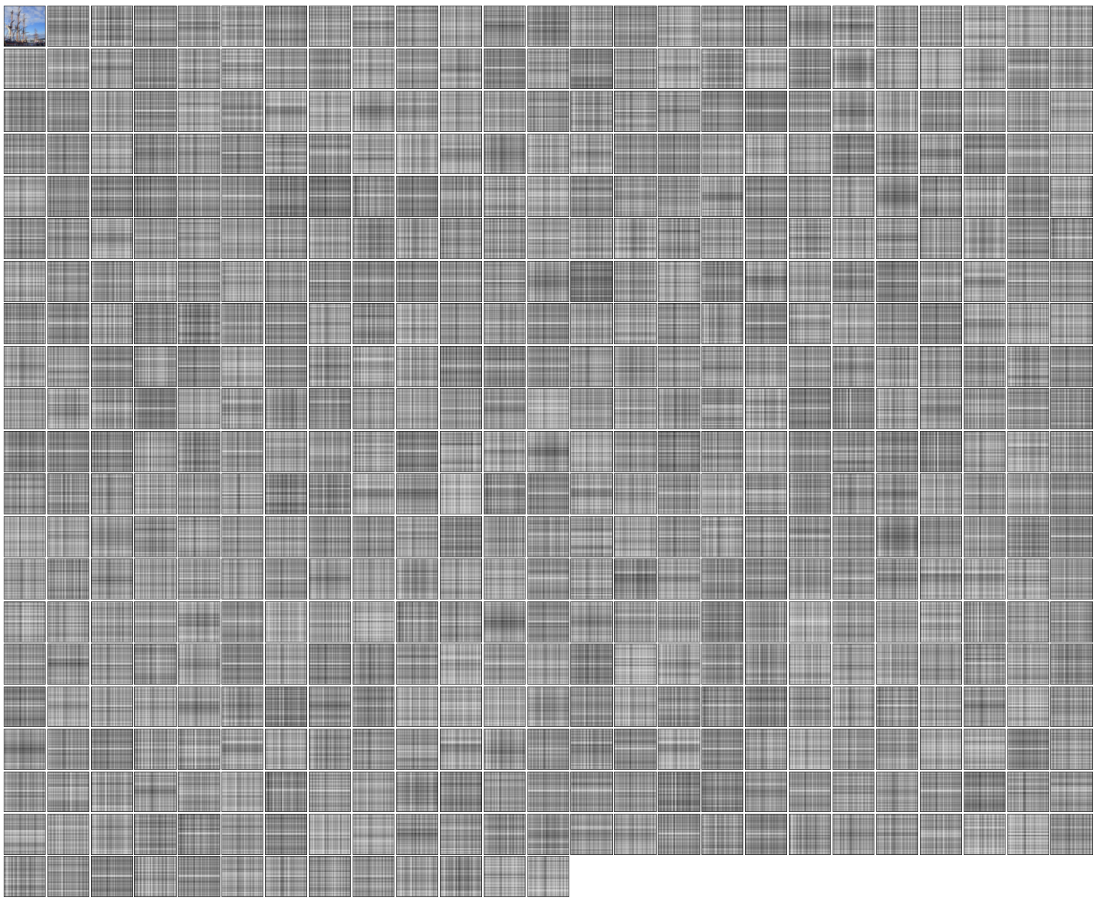

Figure 12. Example attention maps from the **eighth (final) cross-attend** of an ImageNet network trained with **2D Fourier feature** position encodings.

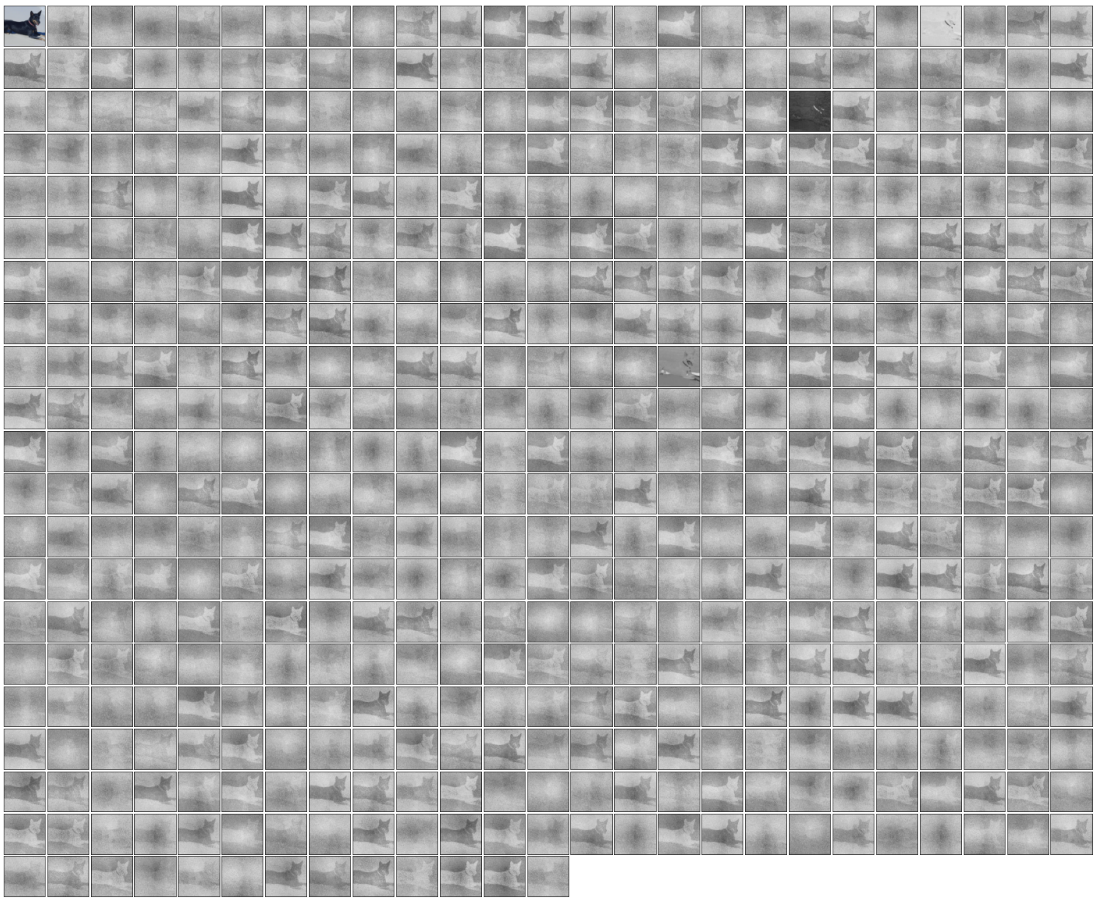

Figure 13. Example attention maps from the **first (only) cross-attention** of an ImageNet network trained with **learned position encodings**.

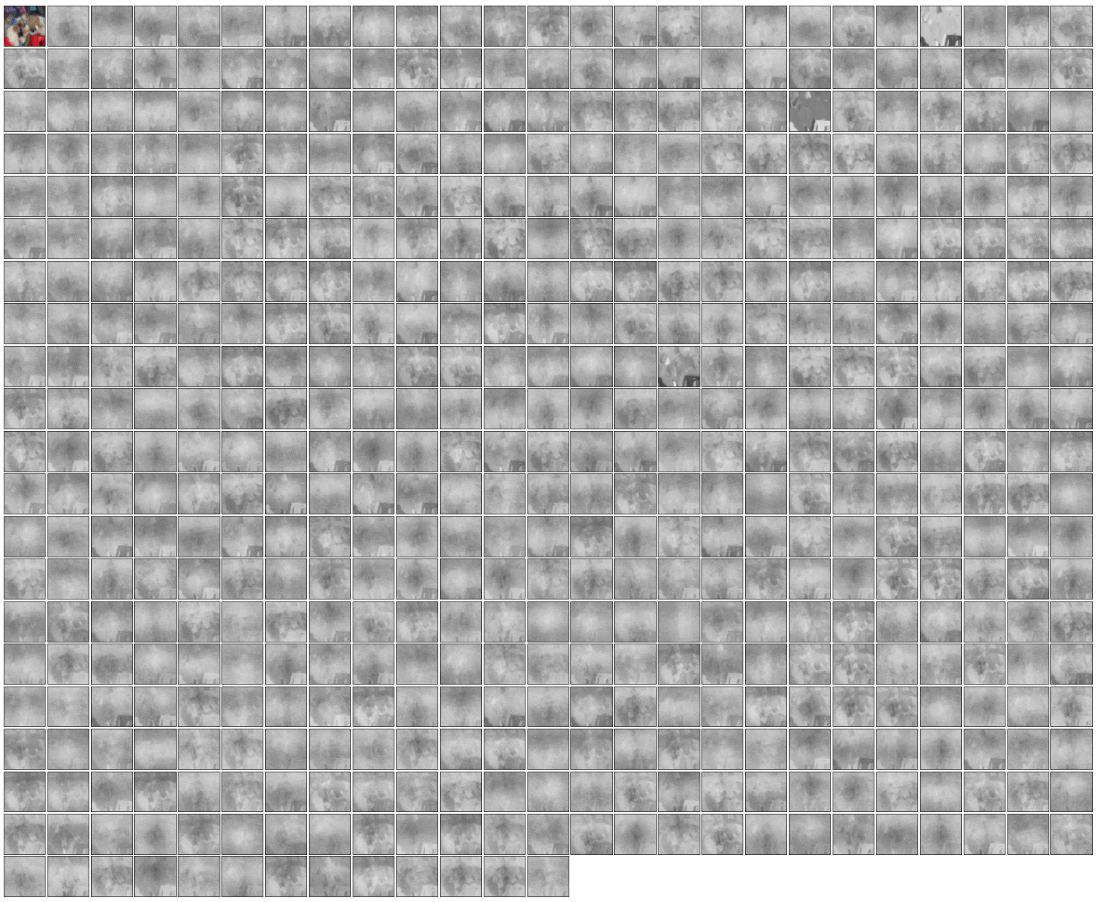

Figure 14. Example attention maps from the **first (only) cross-attention** of an ImageNet network trained with **learned position encodings**.

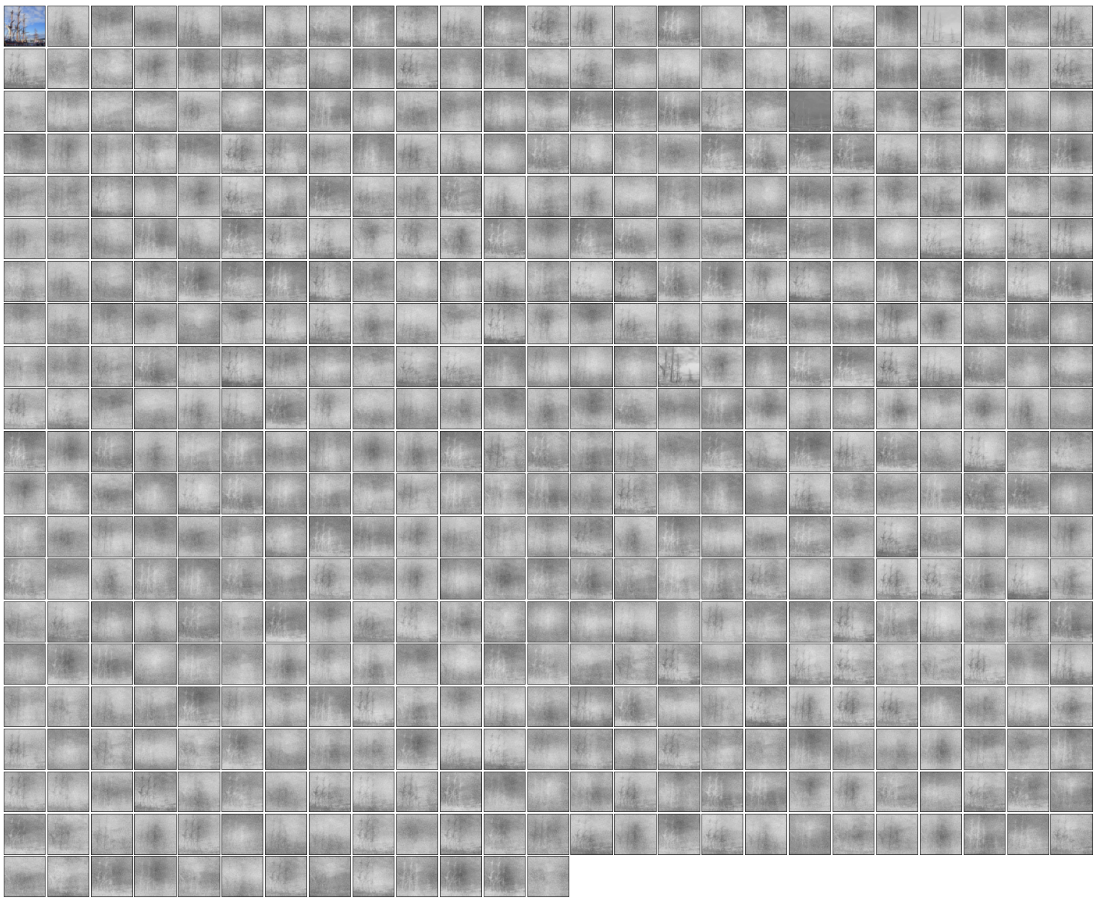

Figure 15. Example attention maps from the **first (only) cross-attention** of an ImageNet network trained with **learned position encodings**.

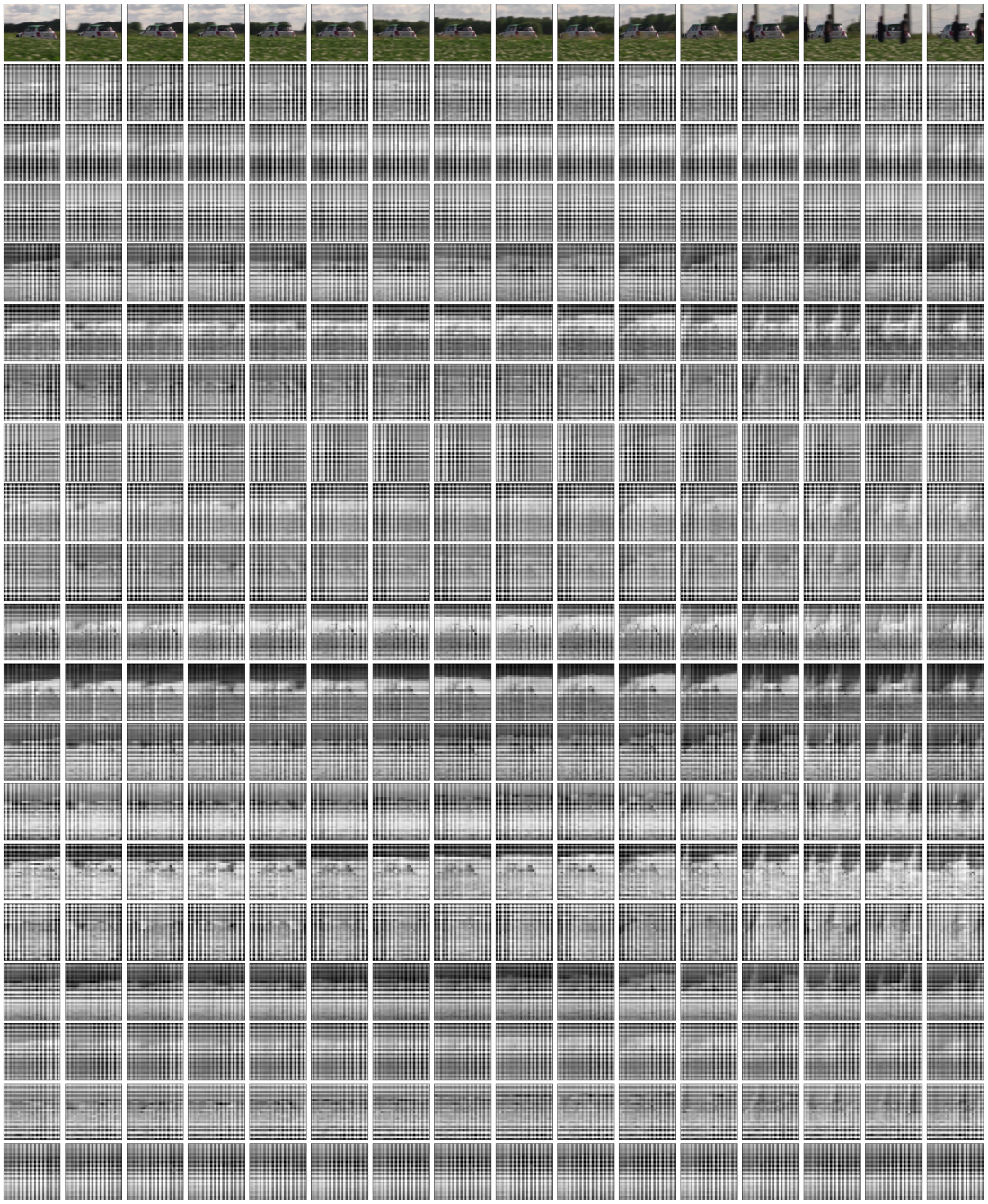

Figure 16. Example attention maps from the **first cross-attend** of an AudioSet network trained on **video only**.

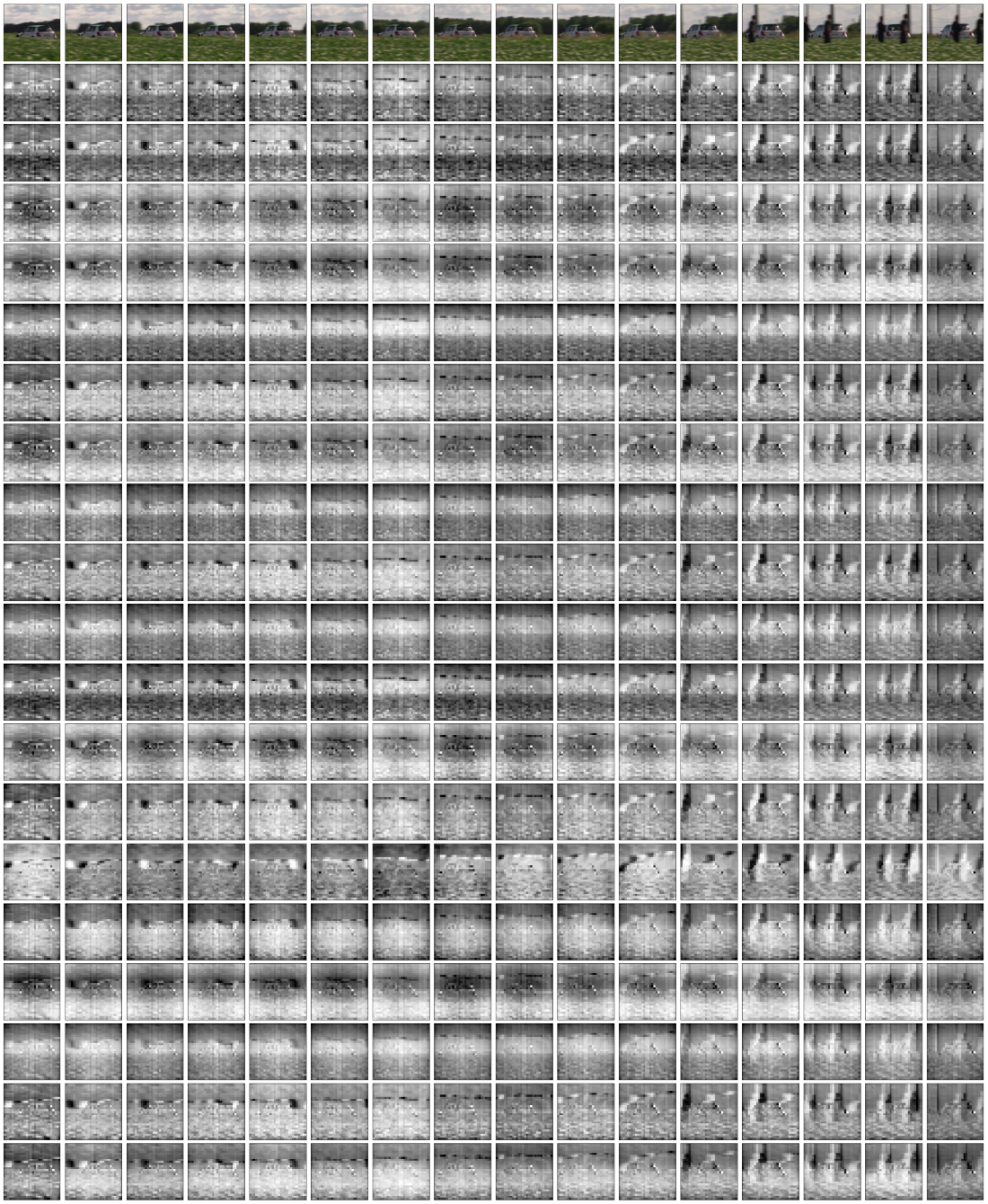

Figure 17. Example attention maps from the **second (final) cross-attend** of an AudioSet network trained on **video only**.

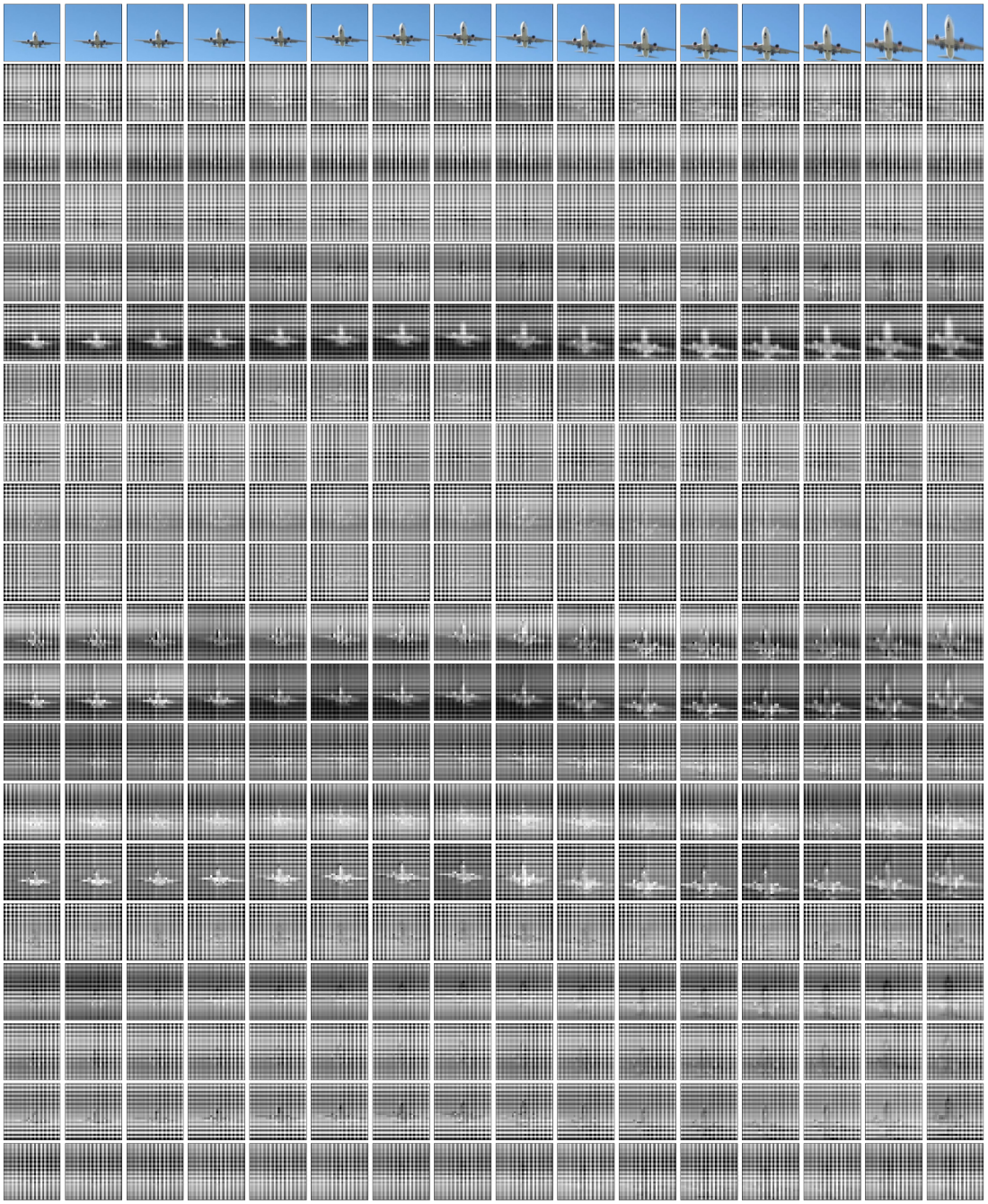

Figure 18. Example attention maps from the **first cross-attend** of an AudioSet network trained on **video only**.