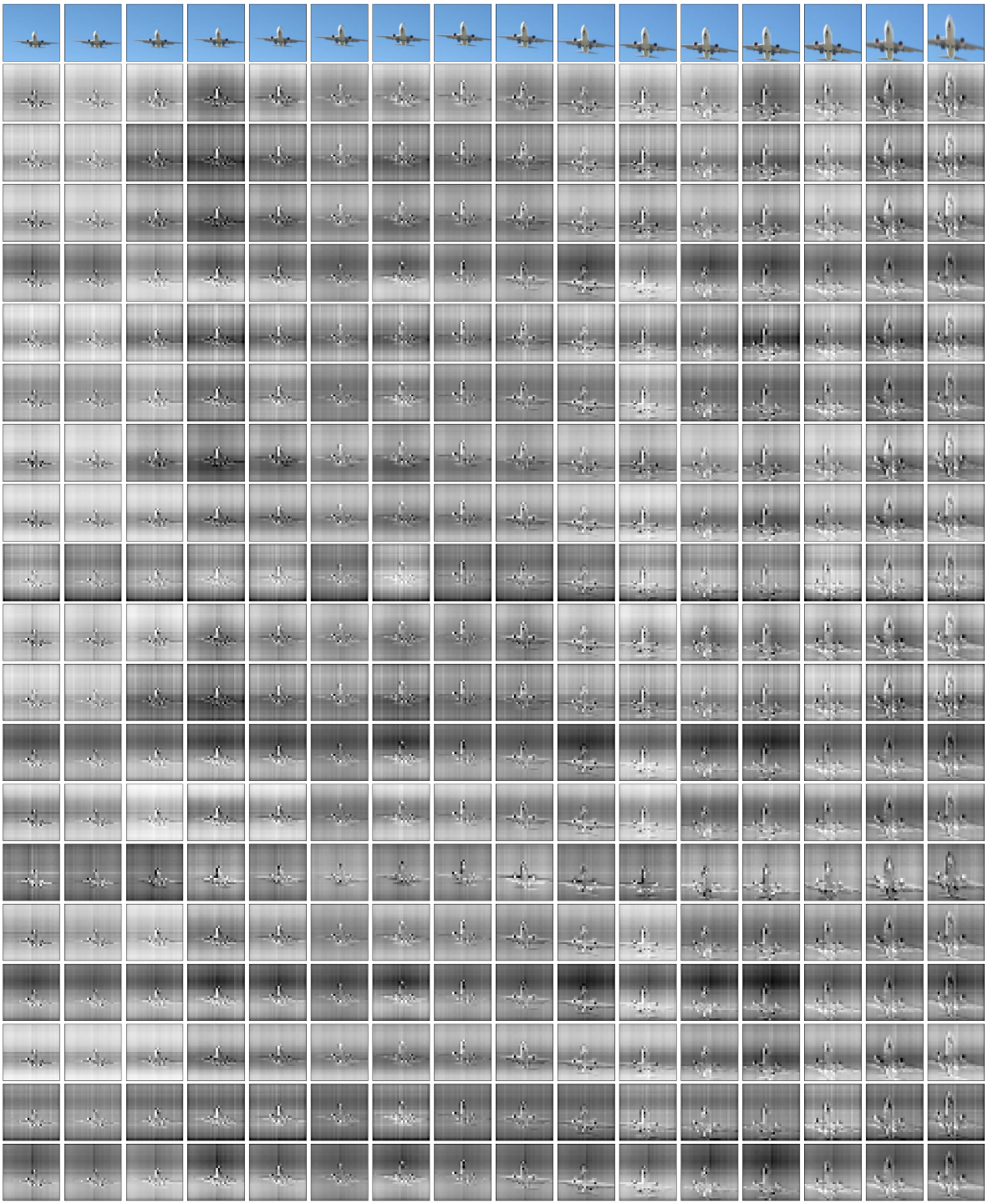

Figure 19. Example attention maps from the **second (final) cross-attend** of an AudioSet network trained on **video only**.

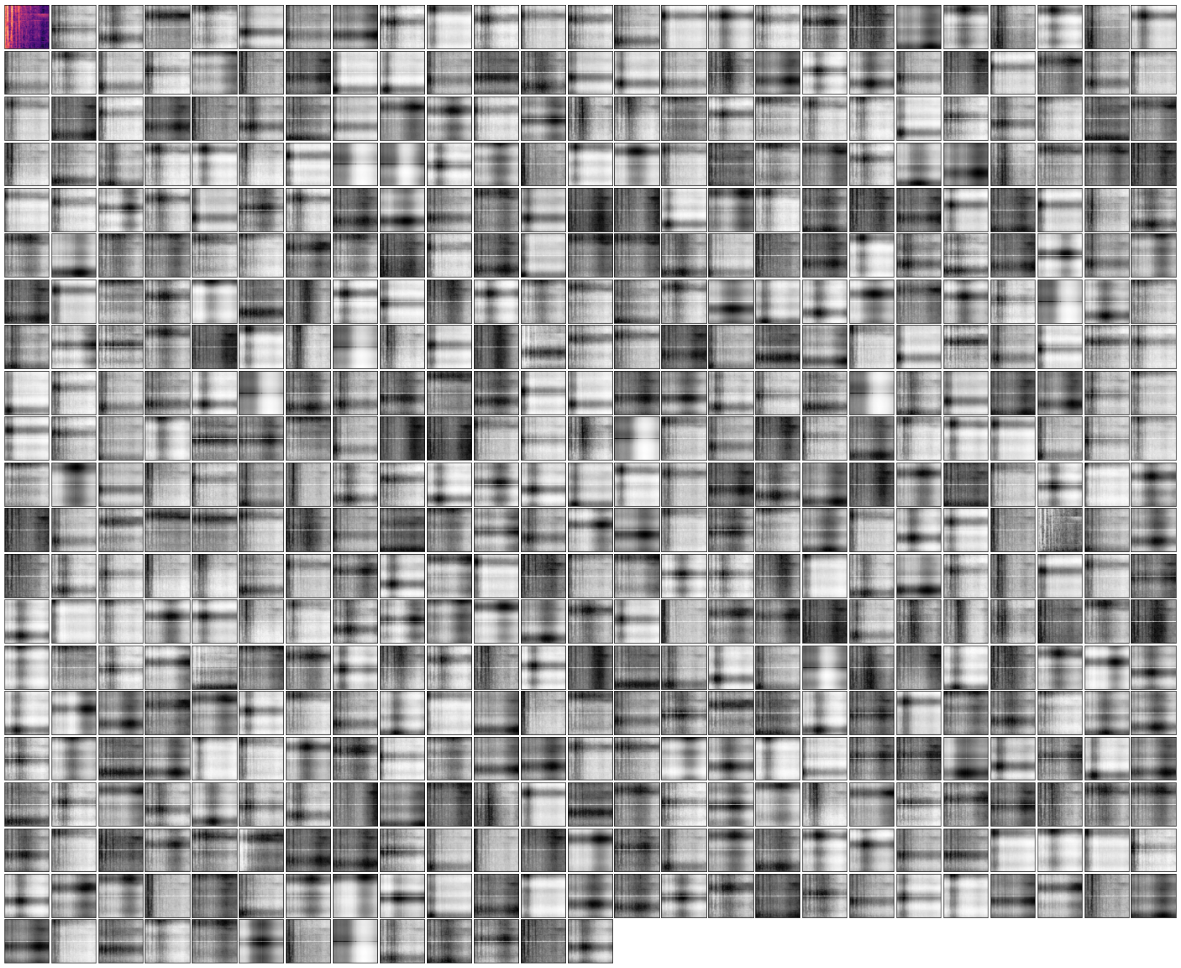

Figure 20. Example attention maps from the **first cross-attend** of an AudioSet network trained on **mel-spectrogram only** (car).

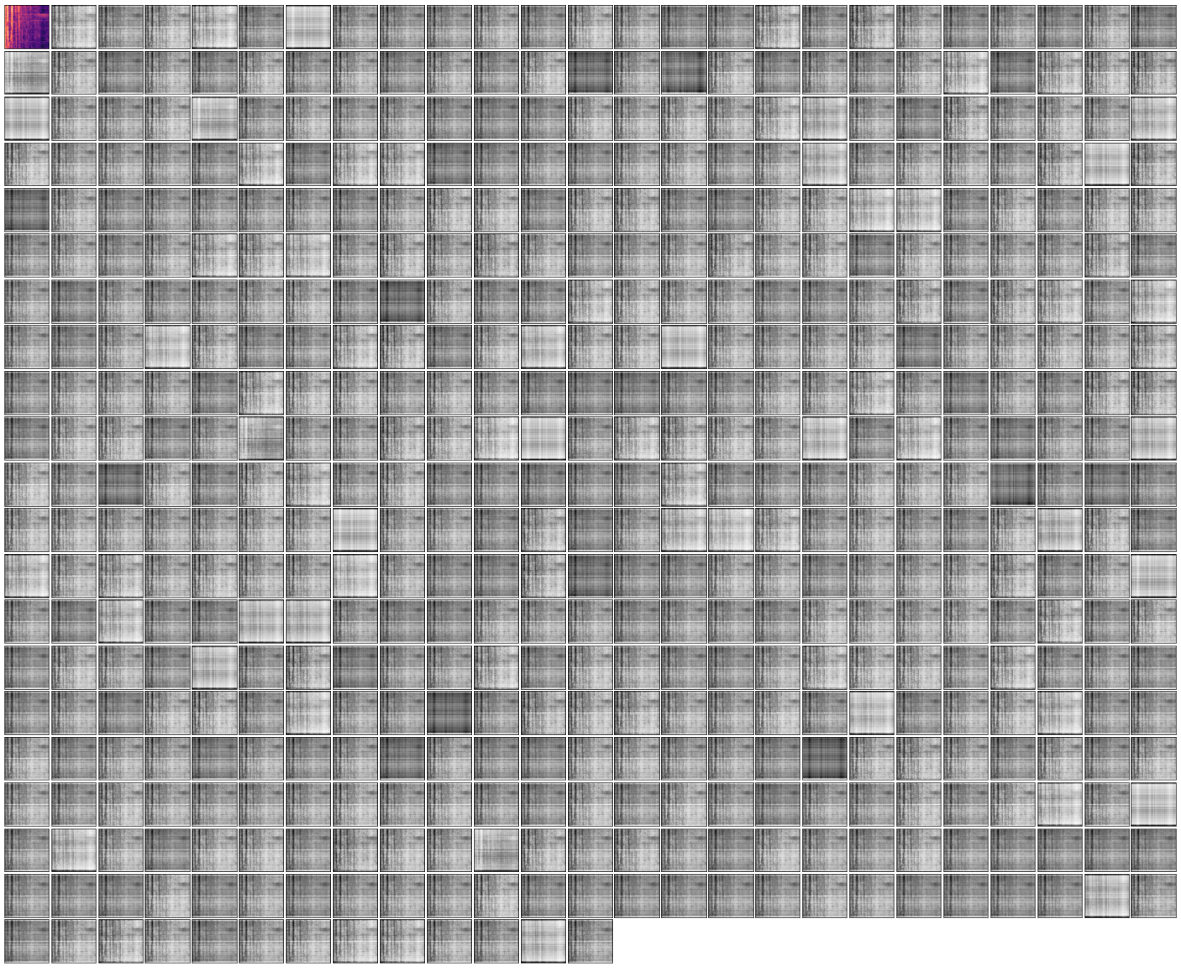

Figure 21. Example attention maps from the **second (final) cross-attend** of an AudioSet network trained on **mel-spectrogram only** (car).

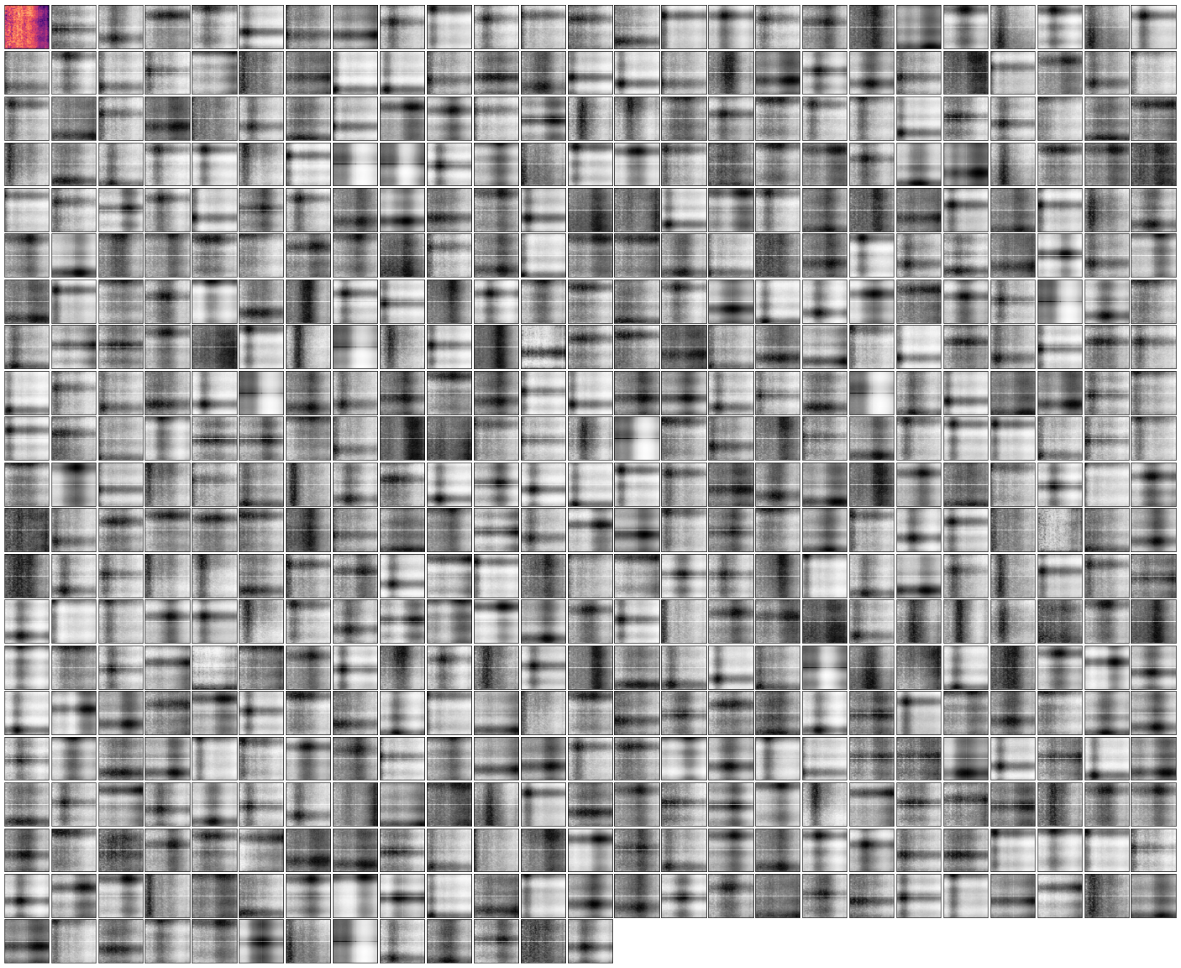

Figure 22. Example attention maps from the **first cross-attend** of an AudioSet network trained on **mel-spectrogram only** (plane).

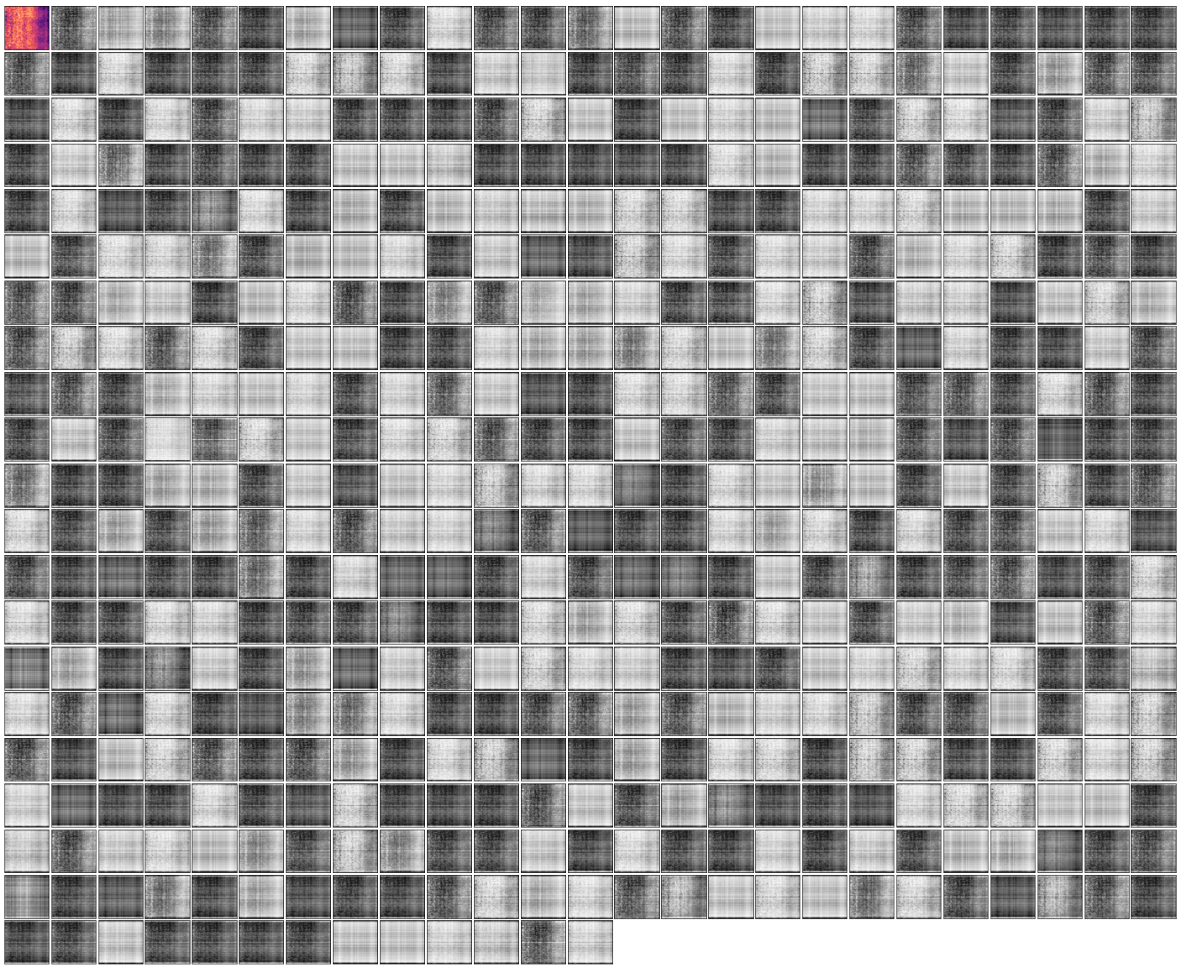

Figure 23. Example attention maps from the **second (final) cross-attend** of an AudioSet network trained on **mel-spectrogram only** (plane).

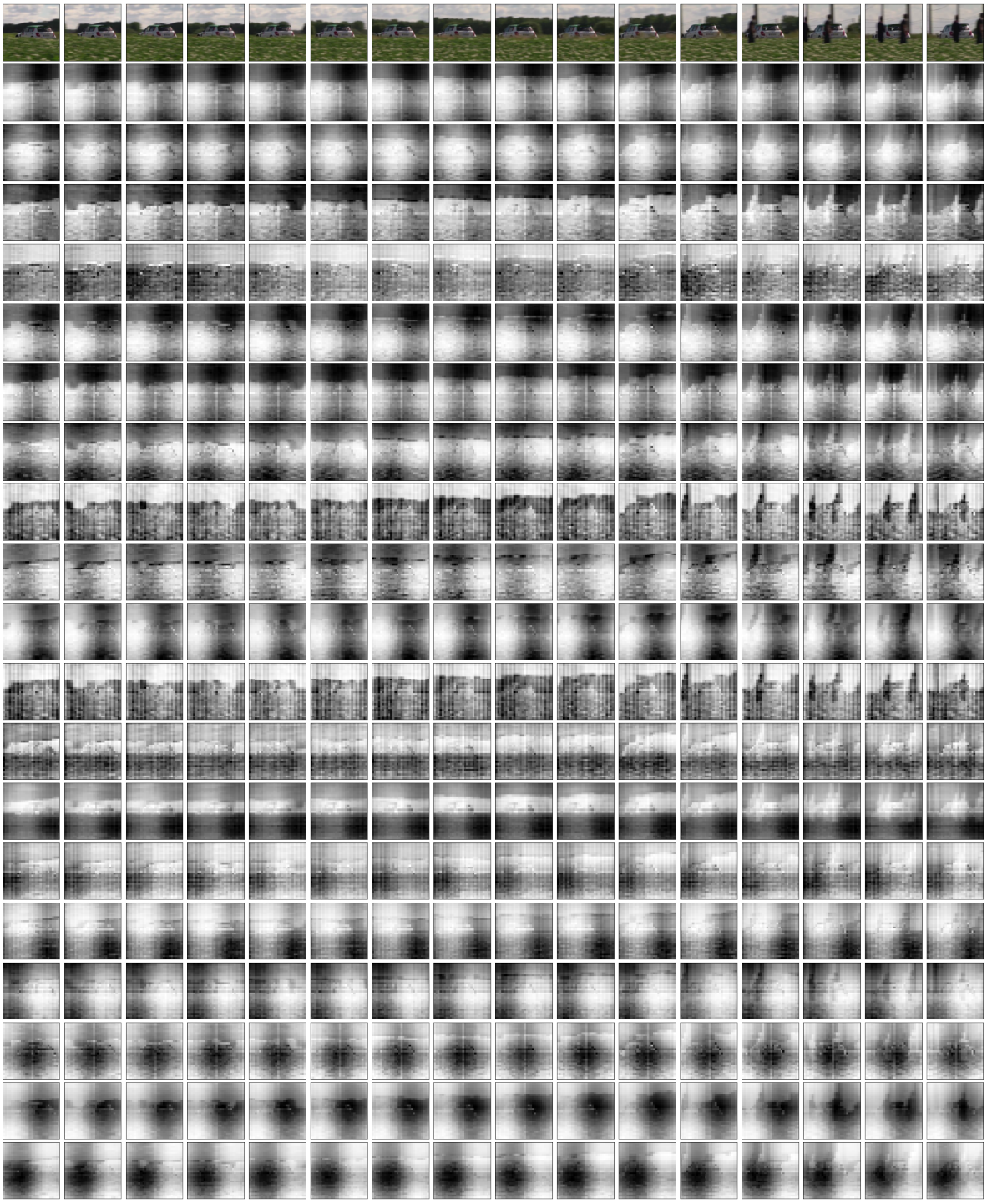

Figure 24. Example attention maps from the **first cross-attend** over the video input subset of an AudioSet network trained on **video and mel-spectrogram**.

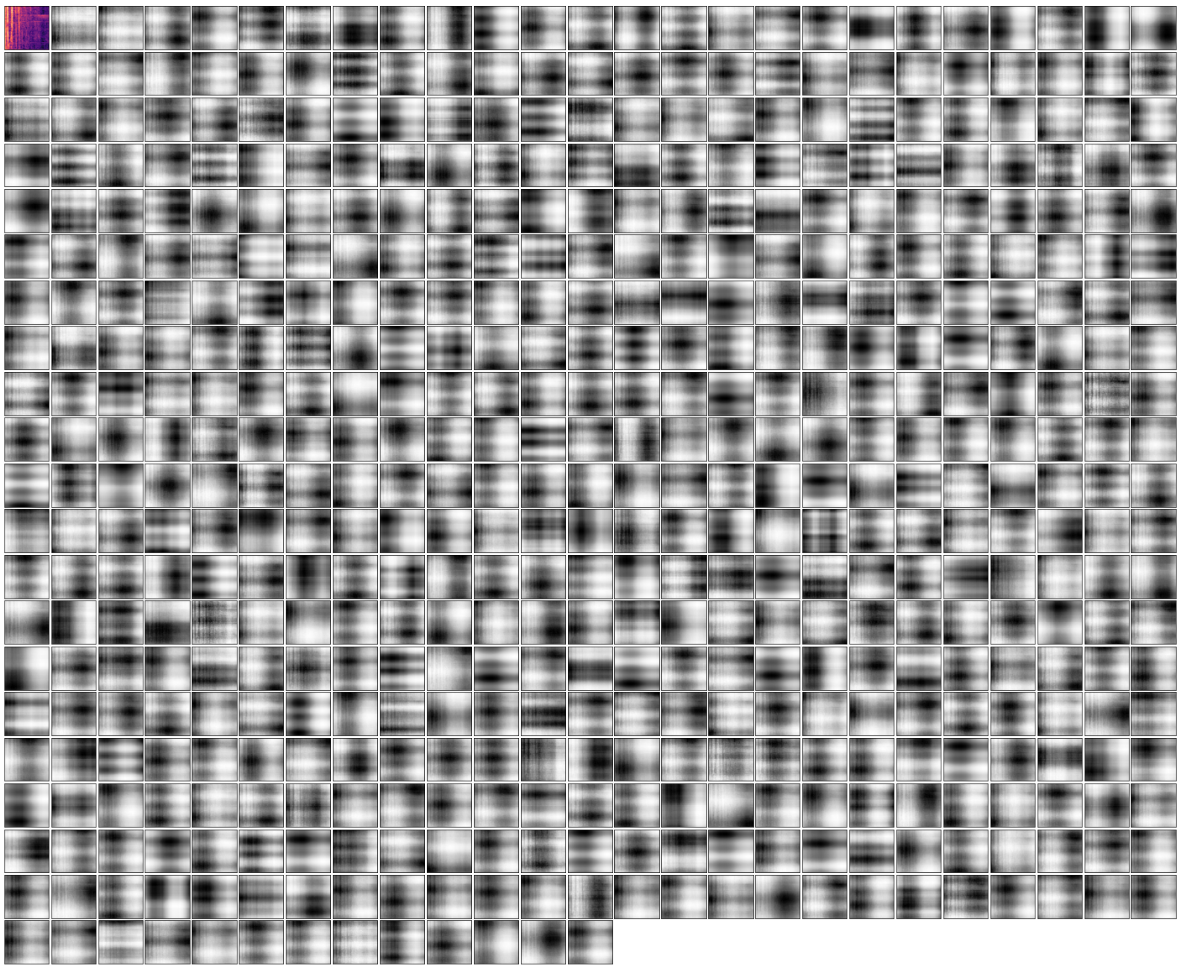

Figure 25. Example attention maps from the **first cross-attend** over the mel-spectrogram input subset of an AudioSet network trained on **video and mel-spectrogram** (car).

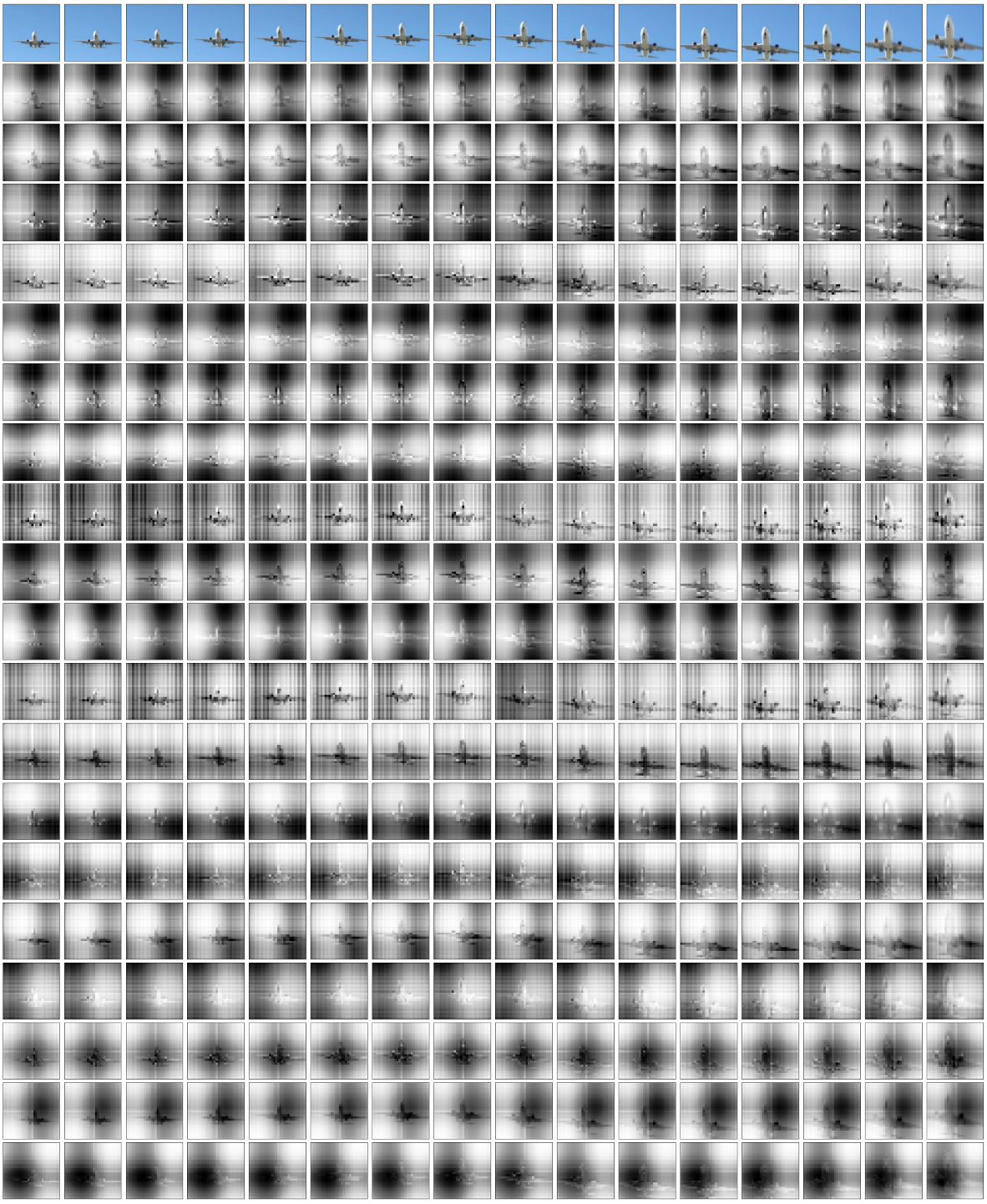

Figure 26. Example attention maps from the **first cross-attend** over the video input subset of an AudioSet network trained on **video and mel-spectrogram**.

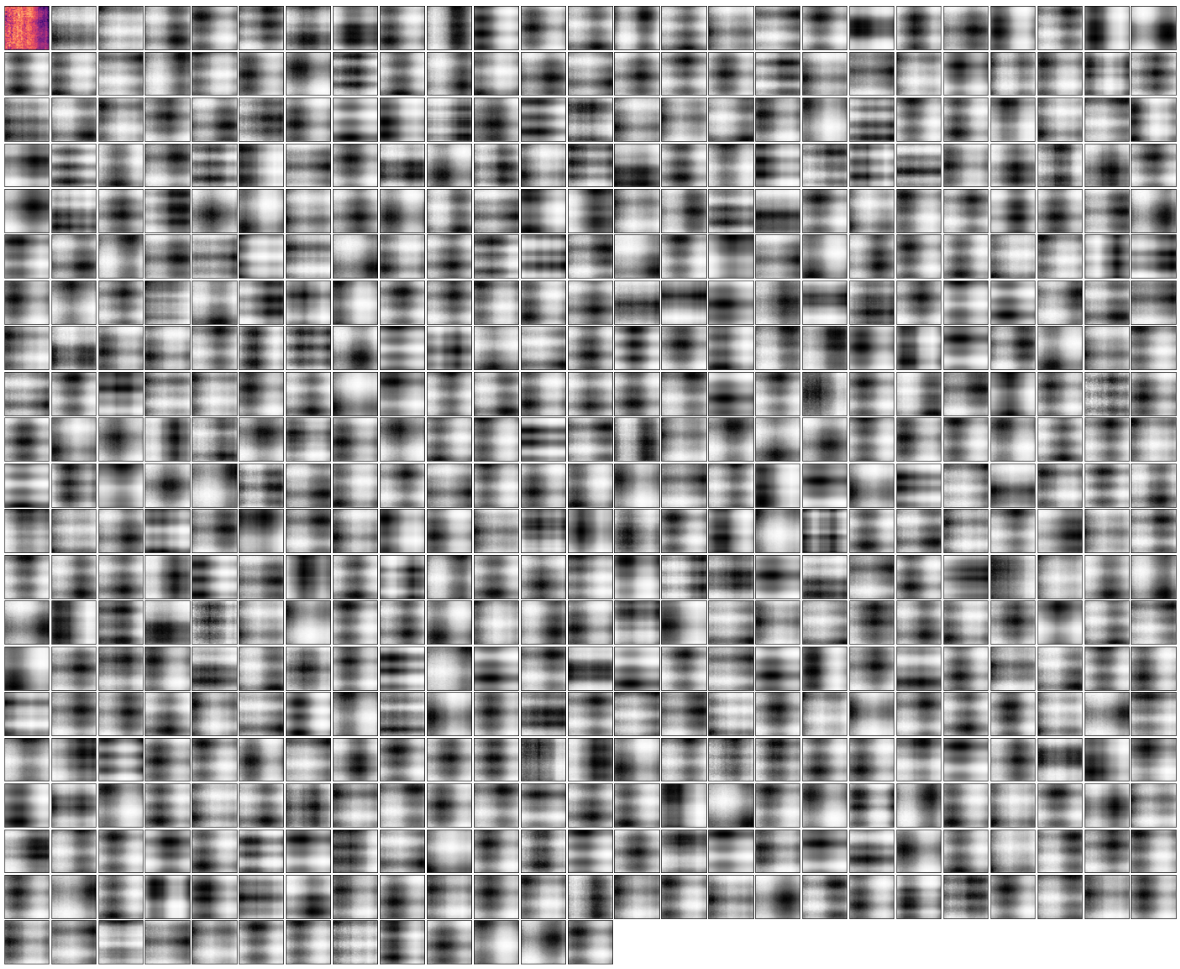

Figure 27. Example attention maps from the **first cross-attend** over the mel-spectrogram input subset of an AudioSet network trained on **video and mel-spectrogram** (plane).

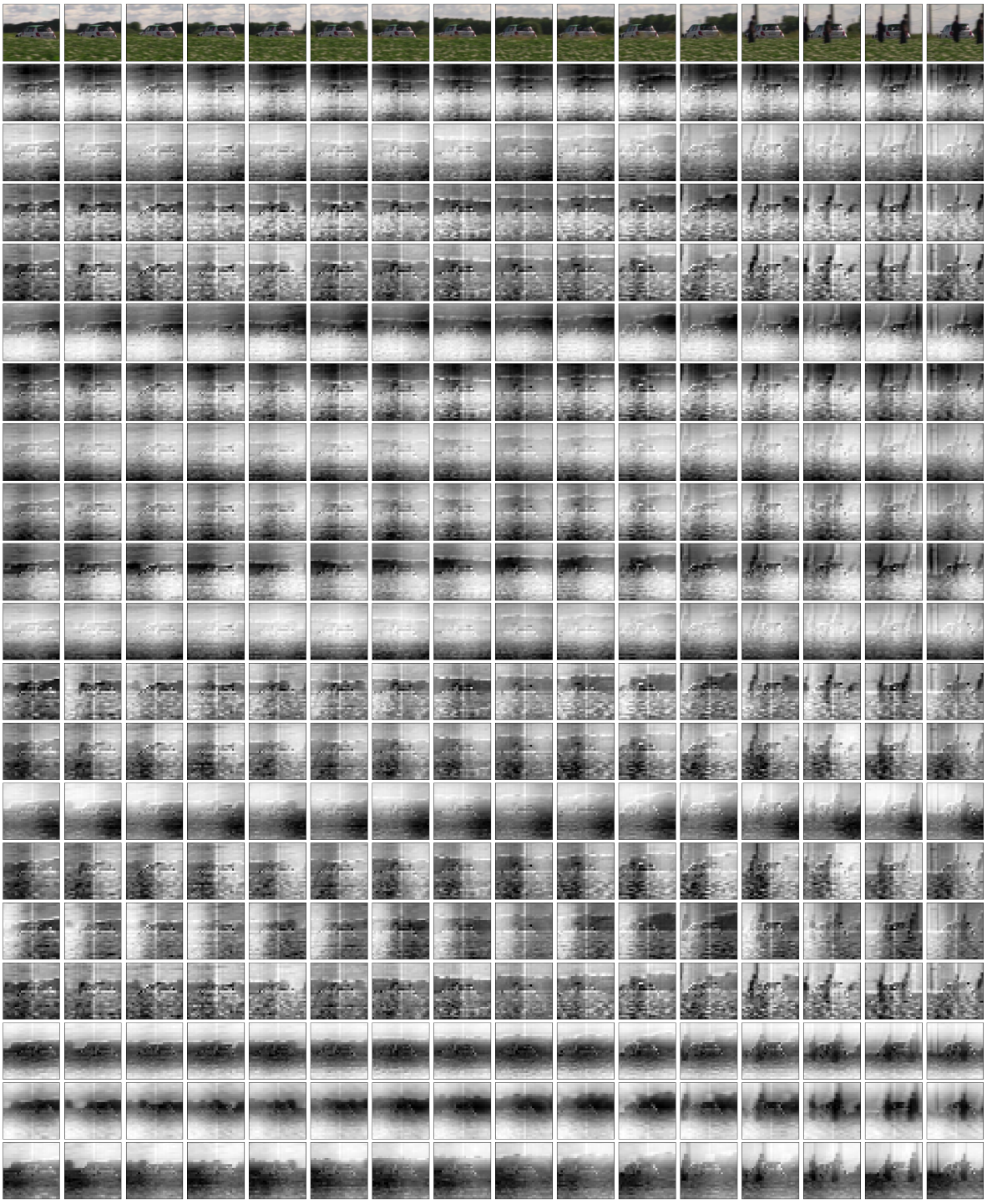

Figure 28. Example attention maps from the **second (final) cross-attend** over the video input subset of an AudioSet network trained on **video and mel-spectrogram**.

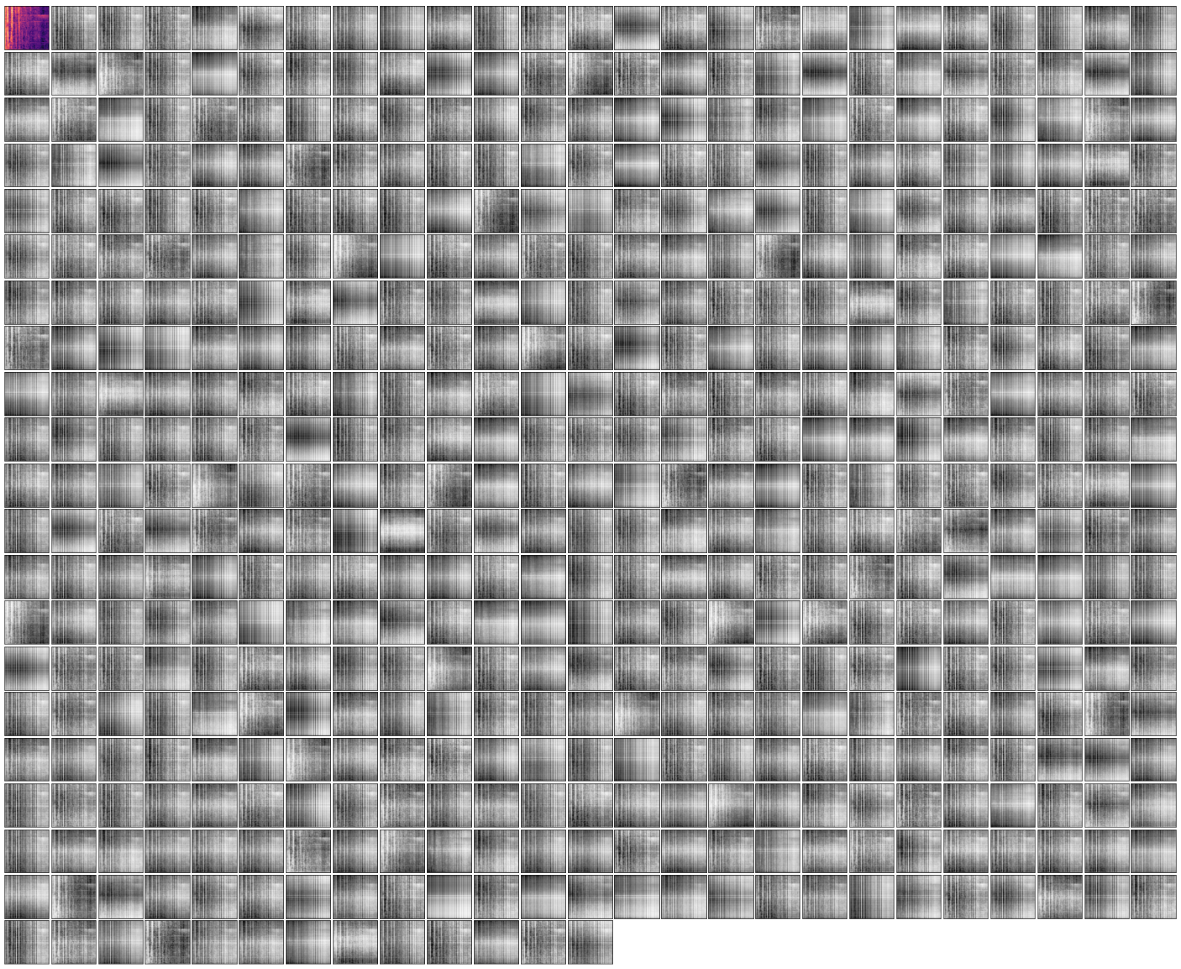

Figure 29. Example attention maps from the **second (final) cross-attend** over the mel-spectrogram input subset of an AudioSet network trained on **video and mel-spectrogram** (car).